

Rochester Institute of Technology

RIT Digital Institutional Repository

Theses

10-2025

Improving the Accessibility of Speech for Deaf and Hard-of-Hearing Individuals Through Affective Captions

Caluã de Lacerda Pataca
cd4610@rit.edu

Follow this and additional works at: <https://repository.rit.edu/theses>

Recommended Citation

de Lacerda Pataca, Caluã, "Improving the Accessibility of Speech for Deaf and Hard-of-Hearing Individuals Through Affective Captions" (2025). Thesis. Rochester Institute of Technology. Accessed from

This Dissertation is brought to you for free and open access by the RIT Libraries. For more information, please contact repository@rit.edu.

Improving the Accessibility of Speech
for Deaf and Hard-of-Hearing Individuals
Through Affective Captions

by

Caluã de Lacerda Pataca

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in Computing and Information Sciences

B. Thomas Golisano College of Computing and
Information Sciences

Rochester Institute of Technology

Rochester, New York

October, 2025

**Improving the Accessibility of Speech
for Deaf and Hard-of-Hearing Individuals
Through Affective Captions**

by
Caluã de Lacerda Pataca

Committee Approval: We, the undersigned committee members, certify that we have advised and/or supervised the candidate on the work described in this dissertation. We further certify that we have reviewed the dissertation manuscript and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computing and Information Sciences.

Dr. Matt Huenerfauth Date
Dissertation Co-Advisor

Dr. Roshan L. Peiris Date
Dissertation Co-Advisor

Dr. Kristen Shinohara Date
Dissertation Committee Member

Dr. Kevin Larson Date
Dissertation Committee Member

Dr. Dan Phillips Date
Dissertation Defense Chairperson

Certified by:

Dr. Pengcheng Shi Date
Ph.D. Program Director, Computing and Information Sciences

Improving the Accessibility of Speech for Deaf and Hard-of-Hearing Individuals Through Affective Captions

by

Caluã de Lacerda Pataca

Submitted to the
B. Thomas Golisano College of Computing and Information Sciences
Ph.D. Program in Computing and Information Sciences
in partial fulfillment of the requirements for the
Doctor of Philosophy Degree
at the Rochester Institute of Technology

Abstract

Captions have traditionally served as a bridge between the spoken word and its written representation, helping make speech accessible to Deaf and Hard-of-Hearing (DHH) individuals. It is worth considering, however, how much from speech is left out by this ‘bridging’ between sound and visuals. This dissertation describes a research project that has, over six studies, looked at this very issue.

We first examined whether there is an issue here at all. What does the experience of DHH individuals with captioning systems tell us about these systems’ shortcomings? For one, we found, captions are felt as monotonous and ambiguous. While communication is multimodal, and DHH individuals also use non-speech cues such as facial expressions or body language to disambiguate a speaker’s intended meanings, these channels are imperfect. Relying on conventional captioning systems is, at their worst, an alienating experience – a lot is lost in these audiovisual translations of speech, and what is lost matters.

Study 2 looked into interventions to captioning systems that could close the gap between spoken words and text. Through various prototypes, it aimed at understanding what dimensions from speech would make captions most helpful: prosody, emotions, or a combination of both. Emotions, we found, were the best compromise between utility and legibility.

Study 3 explored the design space of these ‘affective’ captions. What typographic parameters can best be modulated to depict an emotion’s valence and arousal levels? The investigation looked both at subjective preferences and objective measures. We found that valence should be depicted through color. For

arousal, either font-size or font-weight should be used, with the former preferred for videos with looser legibility requirements.

Studies 4–6 constitute the final phase of this work, looking over how haptics can be shaped to convey a speaker’s arousal, and what are the consequences of doing so. Study 4 experimentally selected a wrist-worn vibrotactile mapping for arousal, identifying a *single short pulse* at 75 Hz (amplitude scaled to arousal) as the best compromise between comfort and discriminability. Study 5 then compared five captioning conditions on longer clips and found that a *combined* approach – valence via color plus arousal via both font-weight and haptics – significantly increased Narrative Engagement for DHH viewers over both a neutral baseline and a visuals-only affective style. Finally, Study 6 measured arousal-decoding accuracy on short clips and showed that adding haptic cues reliably reduced absolute error in perceived arousal, whereas visual weight did not yield a main effect. Together, these studies indicate that multimodal affective captions can be both more *engaging* and more *informative* than conventional captions.

Taken as a whole, this dissertation demonstrates an approach for taking captions beyond verbatim transcription, incorporating affective dimensions of speech. Across six studies, we showed that affective captions are not only technically feasible but also valued by DHH viewers: they can increase engagement, clarify emotional nuance, and support decoding of subtle aspects of speech. By combining visual typography with haptic signals, we offer both conceptual and methodological advances toward more expressive and inclusive captioning systems. Beyond the specific designs and findings, the broader contribution is to reframe captioning as a fertile, multimodal design space capable of accommodating diverse communication needs.

Acknowledgments

I offer my sincere gratitude to Matt, for taking what always seems to me like an impossible leap of faith, and for doing it with such inspiring grace and care; to Roshan, for helping me brave myself outside of my comfort zone and showing that it too can be cool; to Kristen, for serving on my committee, and for doing so with such a keen eye, attentive where mine is blunt, which proved so helpful; to Kevin, for also taking the time and serving on my committee, such a gift for my *subpixel-rendering-rocks* younger self, all the while helping my work stand stronger; to Pengcheng, for helping me understand how all of this fits together in a whole; to all my colleagues at the Center for Accessibility and Inclusion Research (CAIR), some of whom went on to become collaborators and friends, including Abraham, Akther, Andrew, Anisa, Emily, Esther, Garreth, Hidy, Ji Hwan, Laleh, the Matthews, Max, Michelle, Murtaza, Oliver, Paul, Saad, Sarah, Sedeeq, Sidney, Sooyeon, Stephanie, Tom, and Ziming, for being such a brilliant community; to all the participants who took the time to join my many studies, for sharing such thoughtful and exciting perspectives on captioning; to all the students whom I was lucky to supervise, including Andrew, Nathan, Russell, Steve, and Toni, from whom I learned so much; to all collaborators I met along the way, including DJ, Jeremy, Jin, JooYeong, Khai, Lloyd, Mah Noor, SooYeon, and Sue, who continue to show me how exciting all of this can be when we do it together; to my Fulbright cohort, too many to name, with whom I shared, and made sense of, this odd little adventure; to CAPES, the NSF, and the HHS, for all the monies, hopefully put to good use; to Ms. Elbakyan, for still offering help; to Fátima, Fer, Guilherme, and Julio, for taking the time to discuss captions, typography, and design research, when all of this was still rather blurry; to Molly, for helping with my haptic studies, and for being such an all-around nice person; to Paula, for keeping the conversation going; to Claudinho, for planting a seed way back when; to Dan and Patty, and to Patricia and Mike, for helping make this a home away from home; to Abhijeet, Megha, and Nimisha, for such a fortuitous coincidence; to Pamela, for bringing the bread, but not only; to Daniel and Lucas, for all the great work and friendship we've built together; to my co-conspirators Antônio, Justin, Lina, and Sergei, with whom I will (one day) throw off the shackles of oppression; to, of course, anyone I have forgotten, not that big of a surprise given how much age and lack of sleep have eaten me away, but nevertheless a discourtesy, for which I apologize; to my parents, Daniel and Mônica, for always being around, even when 'around' is so, so far away; to my siblings, Caiame and Iara, for helping me navigate this whole growing-old mess; to my kids, Blai and Flora, for all the chaos, so much of it, gosh, but eh, we have our fun; and to my wife, Ana, for everything else.

To Blai, Flora, and the missus.

Contents

CH. 1	Introduction	1
1.1	Part I: Gaps in Current Captioning Approaches	2
1.2	Part II: Affective Captions with Speech-Modulated Typography	3
1.3	Part III: Using Haptic Feedback to Convey a Speaker’s Arousal Levels	4
1.4	Final Remarks	6
PT. 1	Gaps in Current Captioning Approaches	7
CH. 2	Study 1: Deaf and Hard-of-Hearing Folks’ Experiences with Computer-Generated Captions in Remote Meetings	8
2.1	Introduction	8
2.2	Background and Related Work	10
2.2.1	The State of Automatic Speech Recognition	10
2.2.2	Communication Practices Between DHH and Hearing Folks	11
2.3	Study Design	13
2.3.1	Early Designs for Depicting Prosody and Emotion in Captions	14
2.4	Participants and Recruitment	16
2.5	Analysis	16
2.6	Findings	18
2.6.1	Theme 1: Captions’ Dull Ambiguity	18
2.6.2	Theme 2: Communication as an Uphill Battle	19
2.6.3	Theme 3: Reliance on Multimodal Signals	20
2.6.4	Theme 4: Different Contexts Call for Different Solutions	21
2.6.5	Design recommendations from the pilot study	22

2.7	Discussion	23
2.8	Conclusion	24
PT. II	Affective Captions with Speech-Modulated Typography	26
CH. 3	Introduction	27
3.1	Study 2	27
3.2	Study 3	27
CH. 4	Background and Related Work	29
4.1	Paralanguage, and Why Would it Matter?	29
4.1.1	Prosody and the Procedural Encoding of Information	29
4.1.2	Prosody and Emotion	30
4.2	Visualizing a Speaker's Voice	31
4.2.1	The Interplay Between Written and Spoken Language	31
4.2.2	Interventions to go from Expressive Textual Forms to Prosody-Rich Spoken Content	32
4.2.3	Interventions to Go from Prosodic-Rich Spoken Content to Expressive Textual Forms	32
4.2.4	Going Beyond Prosody	33
4.3	Enhancing Caption Text to Improve Viewer's Experience of Captioned Media	34
CH. 5	Study 2: <i>What</i> Dimensions of Speech Should We Depict Through Captions?	37
5.1	Study Design	37
5.1.1	Extracting Prosodic and Emotional Features	39
5.1.2	Caption Design	42
5.2	Participants and Recruitment	45
5.3	Findings	45
5.3.1	Quantitative Data	45
5.3.2	Open-Ended Comments	47
5.4	Discussion	48
5.5	Limitations and Future Work	49
5.6	Conclusion	51
CH. 6	Study 3: <i>How</i> to Depict Emotions in Affective Captions?	52
6.1	Introduction	52

6.2	Phase 1: Evaluating Captions Styles that Depict Valence or Arousal <i>Individually</i>	54
6.2.1	Methods	55
6.2.2	Findings from Phase 1	61
6.3	Phase 2: Evaluating Caption Styles that Depict Valence and Arousal <i>in Combination</i>	67
6.3.1	Findings from Phase 2	67
6.4	Phase 3: Subjective and Objective Performance of the Valence-Arousal Combined Caption Styles Against a Neutral Baseline	71
6.4.1	Methods	71
6.4.2	Findings from Phase 3	75
6.5	Discussion	78
6.5.1	Caption-Style Preferences (RQ3.A & RQ3.B)	78
6.5.2	Factors Influencing Participants' Choices of Caption Styles (RQ3.C)	79
6.5.3	Objective Measures of Performance (RQ3.D)	79
6.5.4	Subjective Measures of Performance (RQ3.E)	80
6.5.5	Design Recommendations	80
6.6	Limitations & Future Work	81
6.7	Conclusion	83
PT. III	Using Haptic Feedback to Convey a Speaker's Arousal Levels	84
CH. 7	Introduction	85
7.1	Study 4	86
7.2	Study 5	86
7.3	Study 6	87
CH. 8	Background and Related Work	88
8.0.1	Interpretation of Haptic Signals	88
8.1	Haptics as Sound/Emotion Translating Channel	89
CH. 9	System Design	91
9.1	Transcription and Emotion Recognition of a Speech Signal	91
9.2	Using a Haptic Signal to Convey a Speaker's Arousal Levels	92
9.3	Typographic Representations of Valence and Arousal Levels	94

CH. 10	Study 4: Using Haptic Feedback to Convey a Speaker's Arousal Levels	95
10.1	Defining the Different Haptic Patterns	95
10.1.1	Rhythm	95
10.1.2	Frequency	96
10.2	Experimental Procedure	98
10.2.1	Analysis plan	100
10.3	Findings from Study 4	100
10.3.1	Haptic pattern rankings	100
10.3.2	Open-Ended Comments	102
10.4	Discussion of Study 4	103
10.5	Limitations	104
CH. 11	Study 5: Visual–Haptic Affective Captions & Engagement	105
11.1	Methods	105
11.1.1	Conditions	105
11.1.2	Stimuli	106
11.1.3	Narrative Engagement	107
11.1.4	Experimental Design	109
11.1.5	Experimental Procedure	109
11.2	Findings from Study 5	111
11.2.1	Narrative Engagement	111
11.2.2	Open-Ended Data	112
11.3	Discussion of Study 5	117
11.3.1	Using Haptic Patterns to Convey Arousal	117
11.3.2	Consideration of Users' Experience with Affective Captions that Employ Haptic Feedback	117
11.3.3	Fine-tuning Color and Font-weight Style Dynamics in Affective Captions	119
11.4	Limitations	120
CH. 12	Study 6: Haptics-decoding Performance	122
12.1	Methods	123
12.1.1	Conditions	125
12.1.2	Stimuli	125
12.1.3	Experimental Design	127
12.1.4	Experimental Procedure	129

12.2	Findings from Study 6	129
12.3	Discussion of Study 6	131
12.3.1	Haptics Helped People Decode Arousal Without Extra Visuals	131
12.4	Limitations	132
CH. 13	Conclusion to Part III	133
CH. 14	Summary and Contributions	135
14.1	Contributions	136
14.2	Publications	138
14.2.1	Additional Publications	139
14.3	Future Directions	140
14.4	Concluding Thoughts	141
	Appendices	167
CH. A	Interview Protocol for Study 1	168
A.1	Introduction	168
A.1.1	Brief Greeting	168
A.1.2	Goal/Purpose of the Study	168
A.1.3	Procedure	168
A.2	The Interview	169
A.2.1	Participants' Experiences	169
A.2.2	New Ideas for Meeting Software	170
A.3	Prototype Demo and Preliminary Evaluation	170
A.4	Collect Demographic Information	171
A.5	Exit the Interview	171
CH. B	12-Item Narrative Engagement Scale	172

List of Figures

2.1	Selection of the prototype designs discussed with the subject matter experts.	15
2.2	Three exploratory designs for an enhanced captioning system shown to DHH interviewees.	17
5.1	The four caption styles used in the test.	39
5.2	Example of the three prosodic features mapped as modulations of three typographic parameters applied to a block of captions.	42
5.3	Font-color representing valence.	44
5.4	The valence color palette under simulation of various types of color vision deficiency.	44
5.5	Font-size being used to represent arousal.	45
5.6	Responses to five Likert-type scales, each row representing one of four caption styles.	46
6.1	Map of the three phases of Study 3, including our final design recommendations.	54
6.2	Diagram of the speech-processing pipeline.	56
6.3	The nine caption styles used in the first evaluation.	58
6.4	Example of one round in our Best-worst scaling setup.	61
6.5	Screenshot of the experiment's platform used for Phase 1.	62
6.6	Charts showing the relative strength and confidence interval for each caption style in relation to valence and arousal, using data collected in Phase 1.	65
6.7	The six caption styles used in Phase 2 of the evaluation.	68
6.8	Relative strength and confidence interval for each caption style tested in Phase 2.	70
6.9	Screenshot of our EmojiGrid implementation.	74
6.10	Box-whisker plots with spread of answers between the five conditions for different Likert scales.	76
9.1	To watch videos, participants would strap the voice coil to their arm, with the device face-down against the inside of their wrist. A laptop would drive both the haptic signals and an external speaker, that played the original sounds coming from the videos.	93

9.2	Example of how typographic attributes can be modulated to convey a speaker’s valence and arousal levels. Here, valence is represented by font-color, with red indicating that the first sentence was said in a negative tone, transitioning to a more neutral and lightly positive tone as they say ‘much caffeine.’ Arousal is shown by changes to font-weight (thickness), reaching its highest when they say ‘it’s fine.’	94
10.1	These charts illustrate how three haptic signal configurations (y-axis, right) respond to changing arousal levels (y-axis, left) over time (x-axis; aligned to word onsets). The dashed-blue lines indicate the predicted arousal values for each word, while the shaded-pink areas show the duration and intensity of each corresponding haptic vibration. The phrase ‘just a hypothetical example’ is spoken with increasing arousal from ‘just’ to ‘hypothetical,’ then decreasing for ‘example.’	97
10.2	Screenshot of the test platform.	99
10.3	Final TrueSkill rankings of the six haptic patterns. LP represents the long pulse; SSP, the single short pulse; MSP, the multiple short pulses. These are combined with two frequencies, 75 Hz and 250 Hz. The skill is shown below each line, with its 95% confidence range shown above.	101
11.1	Screenshot of the page where participants gave feedback about each condition. The image captured from the video was used as a mnemonic device for each caption condition, together with the illustrations and short descriptions. In this example, we see C _{4V+H} (here labeled as ‘caption style 1’), which includes visuals and haptics for arousal, and visuals for valence.	110
11.2	Ridge plot of Narrative Engagement scores across conditions, ranging from 7 to 84. Each ridge represents a condition, with its height indicating the density of scores. Significant pairwise comparisons ($p < 0.05$) between C _{4V+H} and C _{2V} , and C _{4V+H} and C _{1B} , are highlighted by the curly brackets.	112
12.1	Picture of the set-up for Study 6.	124
12.2	Example of the Single Short Pulse (SSP) haptic pattern, originally explored in Study 4.	126
12.3	Example of the font-weight modulation, originally explored in Study 3.	127
12.4	Screenshot of the test platform, showing the captioned video (top) and the arousal rating slider (bottom).	128

List of Tables

6.1	Study 3, Phase 1: Raw and implied results for each one of the 9 styles, applied either for depicting valence or arousal.	63
6.2	Study 3, Phase 2: Raw and implied results for each one of the 6 font style combinations.	69
6.3	Study 3, Phase 3: Distance correlations between participants' valence and arousal measures and the ground truth for each of the five conditions.	77
10.1	The six haptic conditions evaluated in Study 4.	98
10.2	Raw and implied (as per the BWS method) results for each one of the six haptic patterns. In the raw results columns, choosing a pattern 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 pattern was shown in a round but was not marked as best or worst option. The ordering of the table follows the patterns' ascending top-to-bottom TrueSkill values, also shown in Figure 10.3. LP represents the long pulse; SSP, the single short pulse; MSP, the multiple short pulses.	101
11.1	The five conditions presented to participants in this study. The C- abbreviations are used throughout this section. For reference: C _{1B} are conventional captions (the baseline condition); C ₂₋₅ all use font-color to depict valence, with differing approaches for arousal: C _{2V} uses visuals only (font-weight); C _{3H} uses haptic-feedback only; C _{4V+H} uses both visuals and haptic-feedback, and C _{5∅} uses neither, showing only valence.	106
11.2	The five videos used in Study 5.	107
11.3	Median raw scores for each of the four sub-scales and median total Narrative Engagement score. See Figure 11.2 for distribution of scores for each condition. Note that each sub-scale ranges from 3 to 21, and the total scores range from 7 to 84.	111
12.1	Mean absolute deviation from target arousal (0-10). Lower is better.	130

CHAPTER 1

Introduction

Words carry meaning, yes, but nonverbal cues like winks, blushes, body language, and even tone of voice can add depth, nuance, and context to a speaker's message. They will sometimes contradict what's being said, reveal hidden feelings, show changes in mood and confidence, etc. Within this complex, multi-modal dance between words and nonverbal cues, communication takes place.

For Deaf and Hard-of-Hearing (DHH) individuals, the translation of speech to text in captions does not necessarily provide functional equivalency to spoken audio. The process will typically omit meta-speech information such as speaker identification, sentiment, tone, etc, resulting in loss of meaning or confusion [106]. This issue is compounded in automatically generated captions, which are increasingly employed in online communications [116]. Unlike human-generated captions, which can include non-speech information annotations, these automated systems render speech in a uniform, monotonous tone [126], stripping from it acoustic and affective nuances. Their rendering of captions can then lead to communication breakdowns [128] exacerbating, as we will see, feelings of social isolation among DHH individuals. The shift towards remote work since the COVID-19 pandemic [24] has intensified these challenges, highlighting the need for improving captioning technology's fidelity with spoken language.

These omissions are not merely technical shortcomings; they alter how conversations unfold. For instance, captions may fail to indicate *who* is speaking in multi-party calls, leaving participants confused about turn-taking or attribution [128]. Lag in automatic captions can prevent timely responses in fast-paced discussions, while overlapping voices are often transcribed chaotically or not at all [128]. Even when the words are correct, captions strip away tone, sarcasm, or emphasis, flattening speech into text that can obscure intent or emotional nuance [106]. These gaps can cause DHH participants to miss jokes, mistake frustration for seriousness, or feel excluded from the rhythm of group conversation.

In this dissertation, we investigate the challenges posed by this incomplete rendition of some, and not all, elements of spoken speech, and explore approaches that could help mitigate these difficulties. We envision that these approaches could help expand the design space of captioning, with particular benefits to applications that employ automatically generated captions, such as online meetings, live streams, or video calls. These settings have become increasingly prevalent, and given how automatic captions are at times the only means through which DHH individuals can access spoken content, investigating how captions can be made better could bring significant benefits.

The idea of integrating tone of voice into captions is not new [137]. However, and perhaps because automatic captions were historically unreliable, such efforts were limited to bespoke solutions, *e.g.*, giving creators a tool to manually make text more expressive [69, 156], or to help language learners tackle prosody [59]. While automatic captioning remains imperfect [70, 97], the technology has advanced [80] to a point where exploring ways to convey more than just the words through automated approaches now seems feasible. A growing body of research has investigated these possibilities [38, 85, 99, 158, 170, 205], including my own prior work [49, 51, 52]. This dissertation reviews the research I conducted throughout my Ph.D., where I examined pain points in current captioning approaches and the consequences of what they omit, the relative importance of different speech features, and how to represent affective dimensions of speech – first through visuals, then through haptic feedback. Throughout this research, we collaborated with DHH participants to learn not only about their prior experiences but also their impressions of the ideas we proposed and the many prototypes we developed.

This document is divided into three parts. In the first, we investigate the impact of missing paralinguistic cues in captions for DHH individuals. In the second, we explore which paralinguistic features from speech DHH individuals feel captions should convey, and which visual styles are most effective to depict them. In the third and final part, we examine haptic feedback as a channel for conveying arousal in captions and its combination with typographic cues: selecting a comfortable and effective haptic pattern for arousal, evaluating how visuals and haptics affect narrative engagement, and measuring how well participants decode arousal from haptics and/or visuals in a controlled identification task.

1.1 *Part I: Gaps in Current Captioning Approaches*

We start by investigating the state of automatic speech recognition, and the current practices of computer-mediated communication between hearing and DHH folks. Study 1, reported in Chapter 2 (partially pub-

lished in CHI'23 [53]), investigated whether Deaf and Hard-of-Hearing users themselves feel that the gap between captions and spoken speech is significant. In-depth interviews with DHH individuals were conducted to look into their experience with online meetings with hearing colleagues. We asked about what works and what doesn't in current captioning approaches, whether used in professional, educational, or personal settings, and what could be done to alleviate potential issues they highlighted. In so doing, we aimed at answering the following research questions:

RQ1.A In what ways do DHH individuals experience the absence of prosodic and emotional depictions in computer-generated captions, as are used in online meetings with hearing peers?

RQ1.B How can their current experiences and workaround strategies inform the design of new captioning systems that depict prosody and/or emotions?

1.2 *Part II: Affective Captions with Speech-Modulated Typography*

The second part of this dissertation explores the design space of speech-modulated captions – *what* should be depicted, and *how*. Given how Study 1 showed that DHH participants themselves felt that captions are lacking *something*, Study 2, reported in Chapter 5 (partially published in CHI'23 [53]), looked into what exactly this *something* could be. To do so, we designed a set of different captioning styles, each depicting a different combination of features extracted from speech. Comparing these against each other and an additional, conventional caption style that served as a baseline, we sought to answer:

RQ2.A How easily can a speaker's emotions, moods, and emphasis be identified when captions depict prosody and/or emotions in addition to words?

RQ2.B In what settings is the use of these visual depictions of prosody and/or emotion the most appropriate from the point-of-view of DHH individuals?

Having identified in Study 2 the key speech features to represent – specifically, valence (whether a word is spoken in a positive or negative tone) and arousal (whether it is calm or excited) – Study 3 focused on the *how, i.e.*, what visual cues could represent emotions through a caption's typography. For instance, if a speaker is angry, what visual parameters will render their words as having negative valence and excited

arousal? This focus was prompted by a gap in the literature, which provides little guidance on which visual cues are most effective if one wants to typographically represent *emotions*.

We proposed a multi-phase study design, as reported in Chapter 6 (published in CHI'24 [55]), that resembled a 'battle royale.' Beginning with a wide range of visual styles for captions, in each round of the study we ran participant-led evaluations that progressively filtered the number of viable options. The first phase explored how to depict valence and arousal independently. The second phase tested combinations of these styles, further narrowing down the potential styles. Finally, the third phase compared the remaining four styles against each other and a neutral baseline, measuring not only subjective preferences but also how effective each style was in conveying emotions – essentially, whether participants could accurately perceive the emotions being encoded in the typographic modulations. In so doing, we aimed to answer the following research questions:

- RQ3.A Are there caption styles that emerge as preferred by DHH viewers to represent valence or arousal *when depicted individually?*
- RQ3.B Are there caption styles that emerge as preferred by DHH viewers to represent valence or arousal *when depicted in combination?*
- RQ3.C What factors influence DHH viewers' preference for specific caption text styles conveying valence and arousal in speech?
- RQ3.D Do the most preferred methods for conveying valence and arousal in combination, selected in the answering of RQ3.B, outperform a baseline caption text when DHH participants *engage in an emotion-recognition task when watching captioned videos?*
- RQ3.E Do the most preferred methods for conveying valence and arousal in combination, selected in the answering of RQ3.B, outperform a baseline caption text when DHH participants *report on their subjective impressions of how each caption style performs according to the factors outlined in the answering of RQ3.C?*

1.3 *Part III: Using Haptic Feedback to Convey a Speaker's Arousal Levels*

The third part of this dissertation investigates how haptic feedback can complement or extend visual cues in affective captions. Studies 2 and 3 identified valence and arousal as central dimensions to repre-

sent, and pointed to specific visual encodings – font color for valence and either font-weight or font-size for arousal. While these visual cues were effective, participants also raised concerns about legibility, distraction, and the difficulty of finding a universally clear mapping for arousal. In light of these limitations, and building on prior work suggesting that vibrations can aid DHH viewers in perceiving emotional information, we investigated haptics as a complementary channel – extending visual encodings of valence while offering an alternative means of representing arousal.

Study 4, reported in Chapter 10 (partially published in CHI'25 [54]), experimentally compared six haptic patterns created by combining three rhythmic structures with two frequency levels. Using a best-worst scaling method, DHH participants consistently favored the single short pulse rhythm at 75 Hz over alternatives, describing it as both effective and comfortable. In contrast, patterns at 250 Hz or with multiple rapid pulses were often rated as unpleasant or distracting. These findings established the single short pulse at 75 Hz as the preferred encoding of arousal for subsequent studies.

Building on this result, Study 5 (Chapter 11, partially published in CHI'25 [54]) examined whether combining haptic and visual cues could influence narrative engagement when watching emotionally charged video clips. We compared five conditions: conventional captions without affective cues; captions with valence and arousal shown visually; captions with valence visually and arousal through haptics; captions combining both visual and haptic cues; and captions showing valence only. Narrative engagement scores were significantly higher for the combined visual+haptic condition than for both the baseline and visuals-only captions, indicating that integrating haptics enhanced viewers' absorption in the story. Open-ended comments highlighted both benefits (*e.g.*, clearer sense of mood shifts, stronger empathy, heightened presence) and challenges (*e.g.*, distraction when vibrations were too frequent or strong).

Study 6 (Chapter 12) then assessed how well participants could decode arousal levels when conveyed through haptics, visuals, or both. Participants watched short clips spanning the arousal spectrum and rated the perceived intensity of the speaker's emotions. A repeated-measures analysis showed that haptic cues reliably reduced arousal-decoding error, whereas visual font-weight modulations did not yield significant improvements. These findings suggest complementary strengths: haptics improved accuracy in decoding arousal, while visuals, especially when combined with haptics, enhanced engagement over longer narratives.

Together, these three studies address the following questions:

- RQ4 What combination of a rhythmic pattern and frequency, presented as haptic feedback, is perceived as the most effective and comfortable for conveying a speaker's arousal levels, as judged by DHH individuals?
- RQ5 How do haptic feedback and typographic modulations, used alone or in combination, influence arousal depiction and narrative engagement for DHH individuals when compared to a baseline comprised of standard, neutral captions?
- RQ6 How well do modulations of haptics and typography map to how DHH participants perceive a speaker's arousal levels?

1.4 *Final Remarks*

The research presented here addresses a core shortcoming of current captioning technology for DHH audiences: its uniform rendering of speech that omits prosodic and affective nuance. Part I documents the lived consequences of this gap; Part II identifies valence and arousal as priority dimensions and establishes effective visual encodings; and Part III shows that haptics can play a role alongside typography.

Together, these studies reframe captioning as a multimodal design space: visuals (*e.g.*, color for valence, weight for arousal) and haptics (*e.g.*, SSP at 75 Hz) can be composed to improve both *experience* and *information uptake*. We translate these findings into actionable guidance and surface open questions around comprehension and longer-term use (*e.g.*, learning effects, distraction, genre-specific tuning, etc). Beyond the specific contributions, the broader argument is that caption systems can (and should) move beyond verbatim transcription toward expressive, user-tunable, multimodal representations of speech.

PART I

Gaps in Current Captioning Approaches

Includes Study 1

CHAPTER 2

Study 1: Deaf and Hard-of-Hearing Folks’ Experiences with Computer-Generated Captions in Remote Meetings¹

2.1 *Introduction*

Agatha, a participant from this first study, shared how her father would at times say that, since she hadn’t milked the cows on time, their udders were going to explode. Anxiously, she would wonder: ‘Does he want me to call a vet? Are they going to die?’ Eventually, though, it dawned on her that he was joking – something she missed out on at first because the key to getting it was not in *what* he said, but *how* he said it. Agatha² is hard-of-hearing and, she tells us, frequently falls for these pranks because whatever hints of a joke were in her father’s voice are lost by how captions will render his speech deadpan.

Automatic captioning systems have seen remarkable developments in the last decade across many dimensions – they are faster, more accurate, and readily available for little to no cost. Their wide availability has had meaningful implications for the accessibility of speech for Deaf and Hard-of-Hearing (DHH) individuals who may rely on these systems. However, one aspect that has received less attention from the research community is how captions, no matter how accurate, fail to capture the rich expressiveness of spoken language.

Despite ongoing discussions about visual extensions dating from at least the 80s, *e.g.*, [137], captions have changed very little since their inception – for most purposes, they are still rendered as uppercase letters

¹This study was part of a joint project between myself, Matthew Watkins, a graduate student at RIT, Dr. Sooyeon Lee, then a post-doc at RIT, and my co-advisors, Dr. Roshan L. Peiris and Dr. Matt Huenerfauth. I led the study design, stimuli creation, data analysis, and writing of the paper, which was published at the ACM CHI’23 conference [53].

²All names in Study 1 are pseudonyms.

displayed white-over-black in coarse fonts [12].³ The consequences of this flat representation of speech are made worse in settings where automatic captions are employed, for whereas a human captioner might add hints in parentheses of paralinguistic information they consider relevant, automatic captions invariably present words spoken without the added context that non-verbal cues inform.

Human speech, we know, is meaningful beyond just its words. Having captions flatly rendered means ignoring features in a speaker's voice such as prosody (*how loud, melodic, or fast does someone's voice sound?*), vocal quality (*does it sound old or young?*), their disposition (*is it tired or excited?*), or even emotions (*does it carry anger or joy?*) – in other words, conventional captions strip from speech most of its paralinguistic cues.

As motivation and guidance for work in this area, we sought to understand whether this absence of paralinguistic cues in captions negatively impacts DHH users, particularly in communication settings between DHH and hearing individuals. From this stems Study 1, reported in this part. In eight in-depth interviews with DHH individuals, we probed their experience with automatic captions used for meetings with hearing peers, focusing on captions' non-depiction of prosody and emotions, the consequences of this, and workarounds they employ to deal with the challenges that arise in such settings. We sought to answer:

RQ1.A In what ways do DHH individuals experience the absence of prosodic and emotional depictions in computer-generated captions, as are used in online meetings with hearing peers?

RQ1.B How can their current experiences and workaround strategies inform the design of new captioning systems that depict prosody and/or emotions?

Our three main findings are that among other failings, (1) captions' depiction of words is felt as leaving out meaningful dimensions of speech, and (2) in automatically captioned meetings, this can lead to DHH individuals often feeling left out. This state of affairs has been naturalized to the point that (3) some interviewees seem to accept that there are types of conversations that they won't be able to participate in, as if an inherent quality of automatically captioned speech and not a consequence of how these systems have traditionally been designed.

³Interestingly, some captioning standards, *e.g.*, the CEA-708 standard for digital TV, support rich visual features, but common authoring tools are still constrained to more limited, analog-era standards, such as the CEA-608 [115].

2.2 *Background and Related Work*

Automatic speech recognition (ASR) or speech-to-text is a technology that combines methods from computer science and linguistics to convert spoken language into text. It is used in a wide range of applications, from voice-controlled assistants to real-time captioning for DHH individuals. The latter is the focus of the interventions presented in Parts II and III, but to ground their discussion, we will first provide an overview of the current state of ASR technology, acknowledging its progress and highlighting some of its persisting challenges.

Next, we will discuss known issues in settings that employ ASR-tech to facilitate communication, in particular between DHH and hearing folks. Of note, these systems leverage processes where written text is used as a stand-in for spoken audio, a process for which there are still many unresolved aspects, some of which underpin the work discussed in this dissertation.

2.2.1 *The State of Automatic Speech Recognition*

An Automatic Speech Recognition (ASR) system works by converting spoken language into text through a series of computational processes. These systems have seen dramatic improvements in the last 10 years. This is due to massive increases in transcribed data sets, progress in GPUs, and better learning algorithms and model architectures [80]. Such has been this improvement that, for certain transcription tasks, digital systems have surpassed word error rates of human-made transcriptions [70, 80].

This, together with the wide-ranging availability of ASR technology, has commoditized automatic captioning technology – so much so that its use in low-powered portable personal devices has given rise to numerous previously unforeseen applications, such as enabling spontaneous conversations with strangers, supporting employment opportunities (*e.g.*, customer service roles), and assisting with speech training for cochlear implant users [36, 116, 123]. This is significant because it has helped universalize real-time captions as an accessibility option for impromptu or low-resourced settings where DHH and hearing individuals might interact and where services such as sign-language interpreting and CART⁴ might not be available.

⁴Communication Access Real-Time Transcription is a service where a stenographer will provide real-time transcriptions of speech, either remotely or on-site.

Yet, despite its advancements, ASR technology is still far from perfect. Improvements in word error rate (WER) seem to be plateauing [80], while at the same time faith in the WER metric as an end-all correlate to the quality of these systems has faltered. Authors such as Aksënova et al. [3], for instance, argue that WER can be too coarse to describe the *perceived* quality of an ASR system – some words are key to making sense of an utterance; others, not so much. WER may be oblivious to these differences, but users are not.

Compounding this, WER's function as an *average* can hide a system's effectiveness in a long-tail of 'fringe' contexts, with particular relevance for speech accessibility purposes. Factors such as regional language nuances, variations in non-native accents, diverse gender and age groups, latency, domain-specific lexicons, and so on, may negatively impact the recognition performance, and thus also users' overall experience. For example, a study by Feng et al. [63] assessing the capabilities of a cutting-edge ASR system designed for Dutch observed notable variations in performance among speakers of distinct genders, age groups, regional accents, and particularly, non-native accents.

While progress has been made in the recognition of non-standard speech – *e.g.*, Lea et al.'s work with ASR systems and people who stutter [110] or have dysarthria [111] – work is still needed to address the recognition of other non-standard speech patterns. Thus, while recent developments in ASR have improved the accessibility of speech in general, there are still many underserved fringe cases.

2.2.2 *Communication Practices Between DHH and Hearing Folks*

Understanding the limits of current ASR systems is important, but the challenges of computer-mediated communication between DHH and hearing individuals extend beyond them. Recognizing and addressing these broader dimensions is needed when creating inclusive communication environments.

Videoconferencing – *i.e.*, synchronous, multi-party audio-video communication platforms such as Zoom, Microsoft Teams, or Webex – has become ubiquitous in work, education, healthcare, and social contexts, especially since the COVID-19 pandemic accelerated their adoption. It is a setting where these challenges are salient, and can include issues such as the placement of sign language interpreters, the clarity of their signs, and even how the aural-centric nature of the medium shapes its design solutions and constraints [106, 134, 159, 161, 188].

Similarly, when a speaker also employs visuals, *e.g.*, slides, these challenges can compound. The visual channel can easily be overloaded in multimodal communication environments. In a digital lecture, a

hearing person can look at slides while listening to a speaker, for instance, whereas a DHH person may need to switch back and forth between the two, which can lead to visual dispersion and loss of information [39, 109].

Looking at current communication practices between hearing and DHH folks, Elliot et al. [60] found that non-ASR-based communication strategies, whether technologically mediated or not, were felt as being largely unsatisfactory for the latter. This ties with McDonnell et al., who posit that conversation access is fundamentally a social problem, woven into the dynamics and norms between conversation participants. This social dimension predates technology, but is also influenced by it. In particular, they show that assumptions about hearing-centered styles of interaction are often embedded into both technology design and everyday practices, meaning that access depends as much on renegotiating group norms as it does on improving technical accuracy [129].

Illustrating this dynamic, Seita et al. [168] found evidence that hearing speakers will adjust their speech when aware of the presence of an ASR system. They found that, through real-time visual cues, speakers could be 'guided' toward speech patterns that are both more accessible for DHH individuals and easier to recognize by ASR systems [128, 169]. That is not always the case, however: hearing participants sometimes carry misguided assumptions that worsen communication – such as speaking more quickly, overlapping turns, or using unnatural articulation – which have been shown to reduce clarity for both ASR recognition and DHH listeners [168, 169].

2.2.2.1 *Information Overlays and Caption-Design Considerations*

This section sets the stage for later discussions on innovative caption design that can better convey emotional and prosodic cues. Using visual cues as a channel with which to convey additional information beyond the captions themselves is not a simple task. Preferences for caption design vary widely, and while some individuals are open to some more experimental choices, others might resist any changes. The challenge is further complicated by how whatever benefit any new approach offers will have to be balanced with very complex functional requirements, such as considerations of readability in fast-moving text and scene occlusion [46].

Berke et al. [19], for instance, observed a tension in small-group online meetings, with most of their DHH participants' preferences divided into favoring either caption styles with better legibility or those that caused less image occlusion. Similarly, participants in the study by Crabb et al. [46] were divided

between preferring captions placed *outside* of the video area to avoid image occlusion and others who felt this detachment made the watching experience less engaging.

Some researchers have explored overlaying designs of teleconferencing software with meta-information about the meeting or the ASR-system's performance. Amin et al. [10], for example, investigated the importance of identifying the current speaker in a panel discussion, while Berke et al. [20] tested different visual strategies to represent the uncertainty levels that ASR systems assign to each predicted word.

Berke et al. [20] is notable because it uncovered an unintended side effect of conveying information not typically included in captioning systems. In their study, some participants mistakenly understood that captions that displayed their accuracy were *less accurate* than captions that did not, even if accuracy for both conditions was essentially the same. Even though the intervention was successful in conveying new information, this was not enough to update participants' mental model of ASR systems and, thus, did not bring about the expected benefits. This consideration underscores a point present in later studies in this document: understanding information overlays on captions may be an insufficient metric for capturing how these systems influence users' overall experience.

2.3 *Study Design*

The exploration of paralinguistic cues in captions, as highlighted through the example of Agatha's misunderstanding, underscores the impact of non-verbal communication elements on the understanding and interpretation of spoken language, particularly for the DHH community. This study aims at a nuanced view of the experiences of DHH individuals with current captioning technologies. Through in-depth interviews, we sought to not only uncover the immediate functional shortcomings of existing systems, considering their relationship between captions and spoken speech, but to also provide a foundation for developing more inclusive and effective communication tools that recognize and replicate the rich, multifaceted nature of human speech.

We created a semi-structured interview study, with protocol presented in full in Appendix A, which we divided into two parts: (1) an exploration of participants' experiences with remote meetings, including some questions about how speakers' emotions, moods, and other dimensions are handled; (2) new ideas

for how caption design could be enhanced to include these dimensions,⁵ after which we also asked about participants' experiences with and preferences for each.

After the semi-structured section of the interview was over, we showed three early design prototypes of captioning systems that depicted prosody and emotion. The goal here was to gauge participants' initial reactions to each design and to qualify the discussions about how captioning systems can be expanded to include different features extracted from speech.

2.3.1 *Early Designs for Depicting Prosody and Emotion in Captions*

Anticipating that the interviews would suggest a follow-up study, for which we could need to develop captioning prototypes, we decided to use the final few minutes of each interview to show participants videos with some initial designs. While the caption styles used in Study 2 would be redesigned following our analysis of the interviews herein, we wanted to gather some initial reactions that could later inform our work in the subsequent study.

Given the sparse literature on the depiction of prosody and emotions in captioning systems, we consulted with design and typographic professionals to discuss these initial prototypes. This decision was not based on the assumption that they would accurately predict the DHH community's response to these designs – knowledge that is not yet well established. Rather, it was because their experience with type design reflects a quality that is also present in the challenge of 'enhancing' captions: the need to balance maximum expressiveness within very tight functional constraints.⁶ While our goal here was not to establish definite design recommendations for new caption styles, we still needed *some* designs that were sufficiently expressive and to that end, these consultations could be one element helping us find *good enough* design parameters.

In conversations that lasted up to two hours, we showed and discussed an initial set of ideas with four domain experts in type and graphic design. A selection of these is shown in Figure 2.1. We received assorted feedback: using visual particles – colored blobs that floated around the screen following speech

⁵Given the technical nature of a discussion about prosody-based and emotional models to represent speech, participants were given a brief primer on these topics. The circumplex model places emotions in a two-dimensional space (valence and arousal) [160], but for brevity we reduced it to a one-dimensional version: valence only, *i.e.*, words were described as negative, neutral, or positive. The idea of contrasting which words are important was related to prosody, which we also did not break down into its constituent dimensions of loudness, pitch, and duration.

⁶See, for instance, Erik Spiekermann's claim that 95% of a given font has to look like any other font, leaving type designers with only 5% to differentiate their work [90], a setting not too dissimilar to ours.

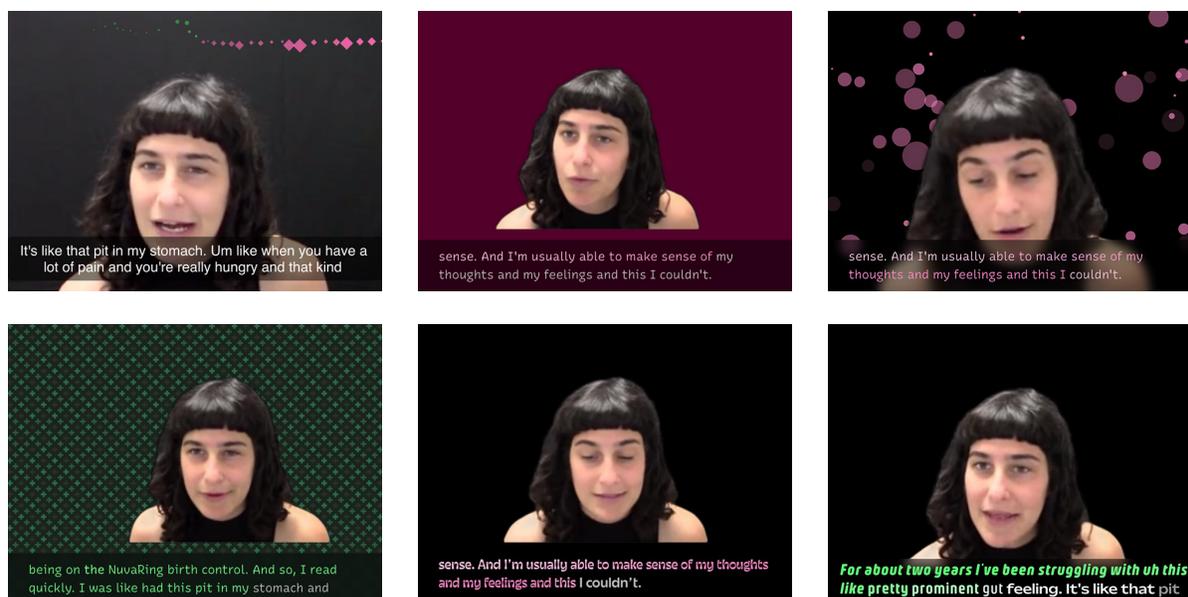


Figure 2.1: Selection of the prototype designs discussed with the subject matter experts.

utterances, as seen in the top left and top right images in Figure 2.1 – as we had done in some of the prototypes, seemed problematic since they would demand too much visual focus from the user, an already overloaded resource.⁷ These experts also suggested that we increase the contrast between various visual properties to make them more apparent. After further refinement based on this round of feedback, we settled on the three designs shown in Figure 2.2 as the most promising to be presented at the end of the interviews in Study 1.

For these designs, we worked with videos available in the Stanford Emotional Narratives Dataset (SEND) [143]. These are short videos of persons looking at the camera and telling stories with strong emotional valence (negative, positive, or both). Included with each file were the video's complete transcription and a set of data-points representing the speaker's self-reported valence levels throughout their speech. This transcription was divided into blocks of 5 seconds, and the valence levels were sampled at 2 Hz.

To approximate the behavior of scrolling captions in the prototypes, we divided the 5-second text blocks into evenly-sized chunks, approximating the timestamps for each individual word. We then assigned

⁷This is specially true for automatic captioning systems, where text is typically added one word at a time (*scrolling captions*), versus *block captions*, where speech is divided into meaningful blocks that are displayed as a whole [132, 152].

a value for the valence of each word, interpolated from the self-reported dataset, and its loudness, measured using the audio from the video and each word's timestamp.

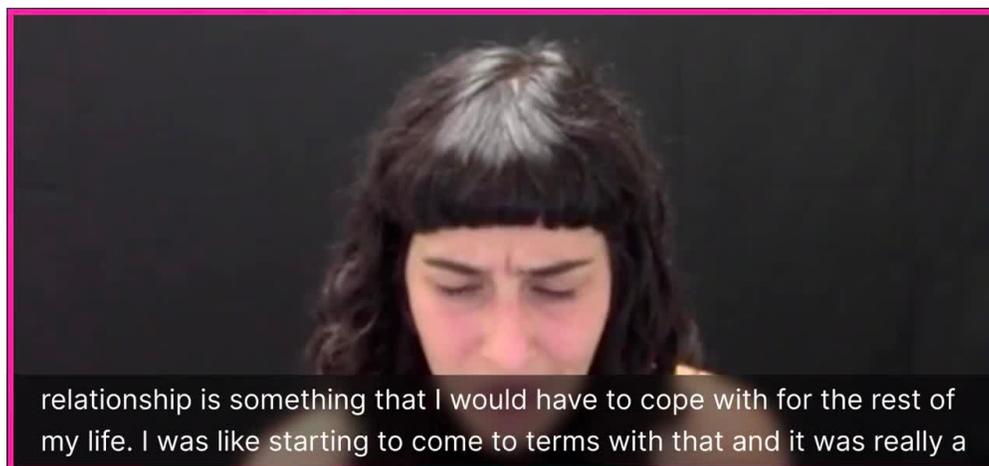
In terms of how valence and loudness were depicted, one design had a border whose thickness varied with loudness and whose color varied with valence (Figure 2.2a); in another, words had their font-weight related to loudness and color to valence (Figure 2.2b); in the last, we used a specially designed 'emotional' typeface comprised of five unique, but related, letter shapes, going from very negative to neutral to very positive (Figure 2.2c); etc. These design decisions were grounded, where possible, in our conversations with experts and informed by the limited guidance from prior work, *e.g.*, [52, 205], though many of the mappings necessarily remained speculative.

2.4 *Participants and Recruitment*

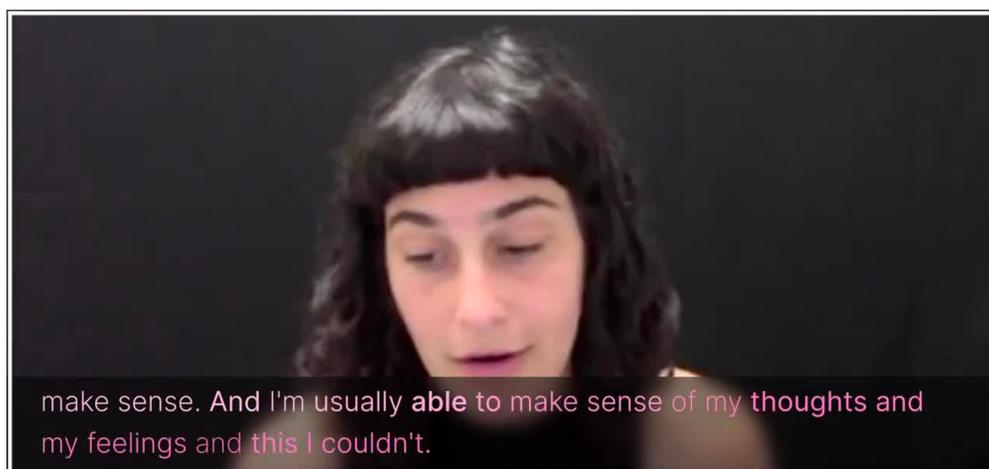
The IRB-approved study ran from March to May 2022. A Deaf research assistant fluent in ASL conducted eight semi-structured interviews with DHH individuals, sometimes in ASL, sometimes in English and ASL, the latter accompanied by an interpreter. Interviews took on average 51 minutes ($\sigma = 14'$). Participants were recruited through social media, particularly through DHH specific Reddit channels, and DHH specific mailing lists. \$40 compensation was offered. The screening factor used was whether the person identified themselves as DHH and had had previous experience working with hearing colleagues sometime in the past five years. Out of the eight participants, five identified themselves as female, two as male, and one as non-binary. Three identified themselves as hard-of-hearing, and five as Deaf. Their average age was 26 ($\sigma = 8$).

2.5 *Analysis*

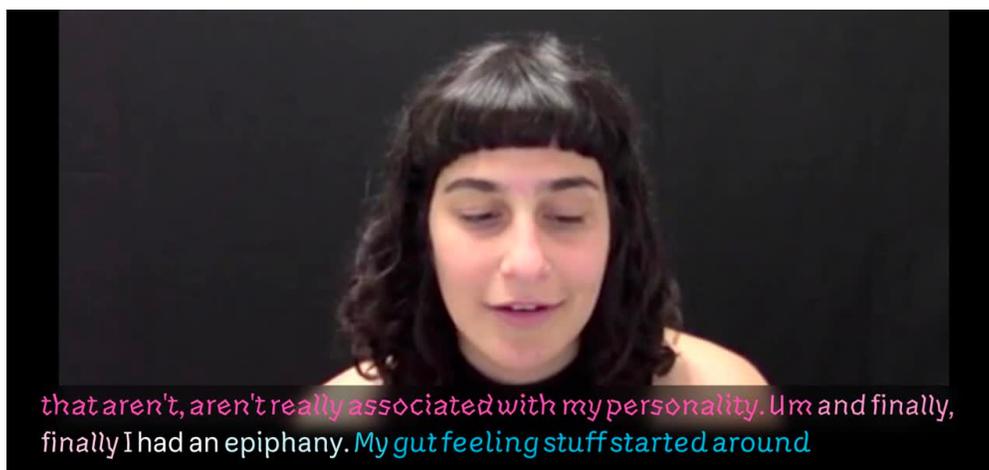
Research assistants fluent in ASL transcribed the eight interviews. On this qualitative data, we performed an iterative thematic analysis [32], taking an inductive, bottom-up approach. Codes were generated directly from the transcripts and refined through several rounds of collating and grouping (*e.g.*, *loss of affective information*), which led to the development of candidate themes (*e.g.*, *Captions' dull ambiguity*). These were reviewed by myself and the team working on this research project in a series of meetings, after which we synthesized a final list of four themes, presented in the next section. A summary of participants' reactions to the prototypes is reported in subsection 2.6.5.



(a) Colored, undulating border



(b) Changes to font-weight and color



(c) Promphan [61]'s valence typography.

Figure 2.2: Three exploratory designs for an enhanced captioning system shown to DHH interviewees.

2.6 Findings

With varying degrees, interviewees related having faced hardships when using automatic captioning systems to communicate with hearing peers. While some issues stem from failings of the current state of automatic speech recognition software, a lack of depiction of prosody and emotions emerged as a cause for *captions' dull ambiguity*. Since interviewees faced this on a nearly daily basis, *communication becomes an uphill battle*, with significant cognitive and emotional tolls.

Strategies to alleviate these ambiguities are diverse and include *reliance on multimodal signals* such as facial expressions, body language, and general engagement. Interviewees indicated that using these cues is not a straightforward, lossless process, and they were therefore favorable towards the promise of captions' depicting prosody and emotions. There was nuance to this preference: given how multifaceted these features can be and given the diversity of what is needed in particular settings, *different contexts call for different solutions*.

Where needed, quotes were edited for clarity and conciseness.

2.6.1 THEME 1: Captions' Dull Ambiguity

Captions' imperfections are felt in different ways. Automatic speech recognition capabilities in live captions have gotten better in recent years, but they still leave a lot to be desired. Agatha: *'Sometimes it's really, really, slow. Someone speaks, and when a few seconds later the caption finally appears the speaker is already on the next topic.'* Eliah: *'Often, when people speak with accents captions will have a lot of mistakes.'* Otto: *'They're horrible, missing context and words. It takes a lot of work to understand exactly what the person is saying.'*

Beyond how latency and imprecision can make the linguistic content hard to understand, there are also consequences related to how a shift in mood can go unnoticed. Alex tells us of a time when there was a quick shift from a casual to a serious topic that was not apparent right away, leading to them *'jumping in at the wrong time and causing my hearing colleagues to look down at me like I'm bad at reading people, which I'm not.'*

Human-made transcriptions of pre-recorded dialogue allow for greater accuracy, but even when spoken words perfectly match written counterparts, still, something seems missing. A common occurrence is

failing to understand whether comments are serious or not. Alex says that since *'comedy relies heavily on tone, hearing people can understand immediately when something is a joke, but my friend, who is also DHH, has a hard time because they're missing that tone.'* Erin tells of a time when *'someone was telling a story that had a specific inside joke, and I had no idea what was going on because it was connected to the tone.'* Otto: *'Sometimes I'll realize it's a joke after they look at me and ask whether I understood it, and I was like "oh, I thought it was serious because the captions seemed serious.''*

Participants complained about the monotonous, droning quality of inexpressive captions. Alex tells that because of this they find it hard to focus on captions: *'I can find it easy to zone out because speech is not really... emphasized?'* Erin finds the contrast between captions and signing hard since she *'grew up accustomed to some use of body language, so it is hard to just watch and read captions all of the time.'* Otto: *'Facial and body language will show a lot of context, while captions are bland.'*

All of this gives captioned speech an unapproachable ambiguity that disproportionately affects DHH individuals. This is particularly true with dimensions of communication that are already inherently ambiguous, such as moods and emotions. Otto thinks that this disconnection is analogous to texting, which *'tends to be devoid of emotion. It's better to interact with the person, to see their real raw emotion, while texting hides it, making it hard to be emotionally transparent.'* For Agatha, when reading captions she tends to miss *'meaning or feeling behind the words.'* To deal with this, she usually has *'to read the full paragraph of what was said, but even then there's a delay in understanding.'*

2.6.2 THEME 2: Communication as an Uphill Battle

Working and studying among hearing peers, our interviewees relate recurrent feelings of isolation. The frequent shortcomings of captioning systems fall almost exclusively on their shoulders, leaving them forced to either speak up or face missing out on what's going on. Ira told us of how in meetings her peers can at times urge her to *'use captioning right away, but I feel awkward because I'm the only person using them. Sometimes I will miss something and feel awkward to ask hearing group mates to repeat themselves; it just feels weird.'*

Sometimes, it's only when they later read a meeting's transcript that what was said becomes clear. Agatha: *'I later understood, but I had to go back and read the transcript to fully understand.'* Eliah: *'It's nice that live captions' transcriptions can be saved as a transcript so I can catch up to what was said.'*

For some, this distance from peers has become naturalized. Erin: *I am curious that I don't know what's happening and I just have to wait there. I know that I am frustrated but at the same time I know that I have to collaborate. I can't expect it to be easy to communicate all of the time.* She later added that *I tend to accept it because of work. Every weekend, they talk about parties and I accept that I am not part of that conversation and just leave it.*

Some environments are more welcoming than others. Otto's manager makes a point of checking how their captions are coming out, saying *"lemme check the captions... Oh no, I didn't mean that," and then repeating themselves until the captions are accurate.* Eliah's boss writes them a summary of what is being discussed because even with an interpreter and captions *there's a lot of overlap and I can't really catch the specifics.*

The flip side is that DHH persons depend on sometimes-lacking goodwill from their hearing colleagues to be included in the conversation. Ada: *Often, my coworkers forget that I need a good environment before I can understand, so they'll be having a conversation with background noise, or not looking at me. I'll still try, but I'll feel alone and left out.* Irene also faces issues with her coworkers' carelessness with how their environment can impact accessibility. When she raises the issue to leadership, they might try to do something, *but the other members of the group are not as willing and, especially since COVID, have reached their limits.* Otto: *I try to be assertive, trying to talk to them, but even if I type in the chat some hearing people don't know how to use it or just ignore it and keep talking anyway. That means I can't do much about it.* When she does intercede, Agatha feels that *with the captions, I'm delayed, so if I had questions I need to ask them to go back on the conversation. I feel like it's annoying to my boss.*

The emotional ambiguity of captions heightened these feelings of isolation. Eliah: *With a large team, it's hard to see their faces and I usually depend on captions. I don't know their emotions, and I feel like I'm not there, not connected with them.* Otto feels missing emotional representation always impacts him: *In general communication, I can't really participate fully. The discussion can be work-related but there's also another discussion that's humorous, and I wouldn't understand. Most of them are laughing and I'm left out, unsure whether they're actually joking or not.*

2.6.3 THEME 3: *Reliance on Multimodal Signals*

Interviewees related being very attuned to how people communicate with their facial expressions, body language, and general engagement. This, some said, is a way to tackle the shortcomings of captioning.

Ira: *'When people are talking I can look and figure out what their thoughts are based on their behavior. With masks, I sometimes miss out on information, so I'll look at their eyebrows or eyes, but it's hard.'* Alex says that to gauge mood or emotion they *'have to look up from the captions at their expressions, body language, and how they react so I can tell what they mean,'* although *'that doesn't mean I capture all the information.'* In describing what makes a speaker's emotions easy to identify, Erin says that it comes out to *'body language; how they're shifting in their seat, how they're moving, their facial expressions, and mouth movement.'*

Cultural differences come into play here. Erin: *'Here in America you can definitely identify it easily, but in other countries, it's challenging.'* Irene: *'It is very hard to understand hearing people's body language and tone, especially through the computer. They tend to sit very relaxed with their hands on their face, or look neutral, while Deaf people are extremely expressive and clear.'*

Technology adds to the complexity of navigating this mosaic of affective signals, and this is present in Agata's comment that *'with captions, sometimes I miss the facial expressions or emotions behind the words.'* The delay in captions makes Ada struggle with trying to listen and read at the same time: *'it's really hard: I have a choice of either listening to the person or reading the captions, but trying both simultaneously takes more work and won't help me.'*

2.6.4 THEME 4: *Different Contexts Call for Different Solutions*

Having introduced to the interviewees the idea that a speaker's tone of voice can serve to signify both different emotions and as a contrastive focus to emphasize certain words, we asked them what they would think was most important to represent in captions. Answers were varied and were tied to what interviewees felt was appropriate for different types of meetings.

Some, such as Otto, claimed that while both dimensions are important, in work environments one should prioritize prosody, *'because I need to understand information better, to pay attention to which word is important. Emotion is important, yes, but I'd rather hold off on that because it's more suitable for general communication.'* Eliah echoed this: *'We don't need to depend on mood because we're here for work. The working environment usually has a lot of discussions so it's important to have emphasis so that we can be involved, discuss more, and ask more questions as deaf people.'*

Others were undecided. Ada, for instance, said that while for her prosody would be generally more important, when their hearing is fatigued, *'I no longer can figure out valence myself, so it would then become*

the more important one.' Alex: *I think both should be included. Valence can show emotion, but not what's important; prosody emphasizes what's important, but not emotion, so how would I know?'*

Others preferred the representation of emotions. In Erin's case, the choice between prosody or valence was almost a tie, but *'emphasis I can figure out, while emotion is really nice to have on the screen so that I know what is going on.'* Ira: *'Emotion is more important since it helps to visualize the full picture, which deaf people usually miss, while emphasis is just for a specific word.'* For Agatha, *'emotions add more depth to words,'* and are thus more important to be visualized.

2.6.5 Design recommendations from the pilot study

Reactions to the design prototypes shown after the interviews were mixed. There was an appreciation of the ideas explored, but not exactly their execution. This issue arose particularly when there was a perceived mismatch between what emotions/emphasis the captions were denoting and what participants were seeing from facial expressions in the video. Eliah: *'The woman on the video was showing distinct facial expressions, there wasn't much change in the border of the first design [Figure 2.2a], but then later on when she wasn't showing much the border became pink or blue.'*

The imprecise alignment between words and sounds was highlighted. Irene: *'I would see the speaker take a breath but there was no break in the captions.'* The display of loudness also seemed misaligned. Ira liked *'the idea in the second design of bolding some words for emphasis, but it didn't seem to match the sentences.'*

Legibility was a major concern, with six out of the eight interviewees having mentioned it. Some of this could be related to the colors used: Erin: *'you get tired of reading, and then the colors start to change, it is confusing to try to understand the tone.'* She also mentioned having some degree of color vision deficiency, which made matters worse. The fonts used were also a source of concern. Ira: *'It was too busy. The font and color changes made it hard to read and look at the person's emotion.'*

Some participants did not notice the border changes in the first design, and some that did found it distracting. Erin: *'The border was awful. Its constant motion would give me a headache.'* Alex: *'Zoom or Microsoft Teams already have a border around whoever is talking, so if you add an additional one tied to the captions it'll be extremely distracting.'* For Eliah, inversely, the border, which reminded him of a similar device used in the video-game 'The Sims,' was functional precisely because it didn't get in the way: *'I liked how the color change represented mood while staying out of the view of the captions.'*

Reactions to the typographic designs (designs two and three) were mixed. Agatha: *I wish the third design [Figure 2.2c] had an easier font to read but I enjoyed the changing fonts because it helped to show the emotions.* Ada: *The best thing about the second design [Figure 2.2b] was maybe the change in font thickness, whether it's thin or thick to show the emphasis, I think that was helpful.* Ira: *Seeing the caption change color was interesting because it helped me separate one sentence from another, while also helping me understand how the person is saying specific phrases.'*

2.7 Discussion

The first goal of the study was to find out in what ways DHH individuals experience the absence of prosodic and emotional depictions in computer-generated captions, as are used in meetings with hearing peers (RQ1.A). Our interviewees discussed the many dimensions in which speech accessibility solutions can fail them. Captions, they pointed out, have many shortcomings. Some come from known limits of current automatic speech recognition systems, which negatively affect DHH individuals' experience of captioning systems [106, 128], and include high-latency and difficulty with non-'standard' accents [3, 63].

Beyond these failings, participants also felt that captions' depictions of words were lacking, leaving out meaningful dimensions of speech. These elements are acoustically present in spoken language but not represented in captions, and their absence creates barriers for DHH individuals. Missing a shift in tone from a serious to a humorous conversation, for instance, was a frequent complaint – and an expected one, given that much humor is conveyed through prosodic markings altogether absent from captions [13].

Our interviewees deal with these challenges in a myriad of ways, but the strategies employed are imperfect. Reading and interpreting text perceived as dull has an additional cognitive toll, and is perceived as a *boring* task. This finding agrees with studies that show that emotional stimuli draw more attention and are better remembered than neutral counterparts [114], an effect that extends to written text [104].

All of these issues leave interviewees feeling as if not part of the group when they are in meetings with their hearing peers. In fact, this is such a common occurrence that for some it has been naturalized as an *inherent* aspect of such meetings, rather than a consequence of how their underlying technologies have been designed.

Our second goal was to understand how these strategies and experiences could inform the design of new captioning systems that depict prosody and/or emotions (RQ1.B). While participants agreed that

including *some* paralinguistic features of speech could help alleviate the ambiguities of ASR-generated captions, they diverged as to which of these dimensions would be most helpful: either emotional cues, prosodic cues, or both.

A follow-up study investigating *how* these captions could look like could thus face a design space too vast to explore. A plausible alternative is to first evaluate *what* non-textual dimensions are most effective to alleviate the communication issues emerging from the interviews, thus allowing future studies a narrower research scope while still measuring whether these expanded captions can help DHH individuals identify paralinguistic dimensions in speech. This is the focus of Study 2, presented in Chapter 5. A 'good enough' design style for captions may be adequate for this first 'what dimensions' study. However, the parameters of this design still require careful consideration. For a detailed discussion on our approach to addressing this in Study 2, see subsection 5.1.2. For a follow-up exploration of how the design space was explored once the 'what dimensions' have been determined, refer to Study 3 in Chapter 6.

In discussing the prototypes shown, responses reflected a diverse set of preferences, allowing some high-level recommendations: (a) Legibility is a notable concern, even when participants felt prosody and emotions were being well represented; (b) Even though participants will generally complement their understanding of captions with a holistic, multimodal understanding – integrating information from such signals as facial expressions and body language – peripheral visual elements used for representing prosody or emotions run the risk of being ignored.

2.8 Conclusion

By examining the experience of DHH individuals with automatically captioned meetings, we found that, in many situations, they feel left behind and unable to fully participate in conversations with their hearing peers. Importantly, we saw that the gap between captions and speech means not only losing access to aesthetic flourishes, but also to key meaning-bearing cues. Losing access to prosodic and affective information led to misread intent, missed topic or stance shifts, and miscalibrated turn-taking – potentially harmful instances of misunderstanding and miscommunication tied to how captions are currently designed to give access to speech.

The consequences extend beyond momentary confusion. Interviewees linked these misunderstandings to concrete, practical harms: hesitation to contribute, breakdowns in coordination during group work,

the need to re-read transcripts after the fact to reconstruct meaning, and cumulative social exclusion when parallel 'humor channels' are inaccessible. Several participants reported normalizing non-participation in specific conversation types, not because the information was inherently inaccessible, but because current caption designs strip speech of the very cues that signal how to interpret it.

Unlike latency or word-error issues though, which are typically framed as technical performance problems, the systematic omission of paralinguistic cues is often treated as if it were an inherent property of captions. Our findings reframe this gap as an *addressable* limitation, motivating our next step.

PART II

Affective Captions with Speech-Modulated Typography

Includes Studies 2 and 3

CHAPTER 3

Introduction

Study 1 showed us that missing out on paralinguistic cues of speech has negative impacts on DHH individuals who employ captions for communication. Using typographic modulations to depict these cues seems like a promising approach to close the gap between spoken and captioned speech. However, it was not clear from the interviews which cues would be most beneficial to depict, and how to depict them in a way that is both effective and acceptable to DHH individuals.

3.1 *Study 2*

Part II of this dissertation aims to address these questions. In Study 2, presented in Chapter 5, we investigated whether DHH folks derive greater benefit from captions that depict prosody, emotions, or a combination of both. As such, we aimed to answer the following research questions:

RQ2.A How easily can a speaker's emotions, moods, and emphasis be identified when captions depict prosody and/or emotions in addition to words?

RQ2.B In what settings is the use of these visual depictions of prosody and/or emotion the most appropriate from the point-of-view of DHH individuals?

3.2 *Study 3*

Having found strong evidence that captions that show only emotions are a good balance between utility and legibility, Study 3 (Chapter 6) investigates the design space of such captions. While there are

many studies exploring how to depict prosodic features through typography, little has been done as to how to depict the combination of valence and arousal. As such, Study 3 aims to answer the following research questions:

RQ3.A Are there caption styles that emerge as preferred by DHH viewers to represent valence or arousal *when depicted individually?*

RQ3.B Are there caption styles that emerge as preferred by DHH viewers to represent valence or arousal *when depicted in combination?*

We divided the third 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 RQ3.A can help us narrow down viable styles for each aspect before combining them to address RQ3.B. In order to enrich our understanding of DHH individuals' preferences and the reasons for their choices, we also ask:

RQ3.C 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.D Do the most preferred methods for conveying valence and arousal in combination, selected in the answering of RQ3.B, outperform a baseline caption text when DHH participants *engage in an emotion-recognition task when watching captioned videos?*

RQ3.E Do the most preferred methods for conveying valence and arousal in combination, selected in the answering of RQ3.B, outperform a baseline caption text when DHH participants *report on their subjective impressions of how each caption style performs according to the factors outlined in the answering of RQ3.C?*

CHAPTER 4

Background and Related Work

This chapter provides a panorama of concepts and related work that ground Studies 2 and 3. It is divided into three sections: First, we give an overview of how paralinguistic cues help people make sense of speech. Second, we discuss past research that has looked into closing the gap between spoken and written language, highlighting typographic approaches. Finally, we review previous studies that have explored how captioned text can be enhanced to improve viewers' experience with captioned media.

4.1 *Paralanguage, and Why Would it Matter?*

4.1.1 *Prosody and the Procedural Encoding of Information*

If we are to intervene in captioning systems so that they more accurately depict spoken language, it is important to understand how speech has its meaning influenced by how a speaker modulates their voice. A good starting point is prosody, which shapes much of what we perceive as tone of voice [14]. When speaking, a person will articulate prosodic variables such as pitch, length of sounds, loudness, timbre, etc. In so doing, speakers are encoding procedural information: unlike words, which can be directly mapped to concepts – *e.g.*, the word *cat* is related to the feline animal – procedural encoding in prosody helps guide a listener through a sentence's *plausible* meanings [202].

Said differently, at the outset, a given sequence of words might have multiple incongruent, but plausible, interpretations. Through prosody, a listener may narrow their search space of what interpretation is most closely tied to a speaker's intended meaning. For example, in the sentence 'that dog loves you,' we can imagine different interpretations depending on what word is emphasized by the speaker: If they stress *that*, they may be contrasting this one particular dog with another; if *you*, maybe that same dog

has different feelings towards someone else; etc. All of these meanings are initially plausible when one only considers words. It is through prosody (along other disambiguating channels) that a listener is able to eliminate implausible interpretations.

4.1.2 *Prosody and Emotion*

Exploring this interchange between how meaning is shaped by prosody, some studies have experimented asking individuals to rate what emotion they perceived in different sound excerpts, some of which were spoken in the participants' native language, some of which were not. Semantic content is predictably the main factor determining the affective label assigned by participants – a tragic story will likely remain tragic if the speaker utters it in a jolly tone – but even if the semantic content is removed, like when the speech is in an unknown foreign language, emotional patterns can still be apprehended [47, 102]. This suggests that prosody is a powerful channel for conveying emotions, and that it can be understood at least partially across linguistic and cultural boundaries.

In a similar vein, de Moraes and Rilliard [57] found meaningful correlations between specific patterns of change in pitch and duration and sentence mode and emotions. Sentence mode (*i.e.*, whether the sentence is a statement, question, command, or exclamation) aligned with the overall intonation contour (*i.e.*, the shape of pitch levels as we move across a sentence), whereas emotions were associated with average intonation values. For example, regardless of emotion, a question sentence typically has a peak and a pronounced drop in pitch between the second-to-last and last syllable. In contrast, a sad intonation generally features a lower average pitch compared to a joyful one. In other words, prosody was found to convey two distinct dimensions of meaning, namely, sentence mode and speaker emotion. The point here, then, is that while part of this prosodic dimension of meaning is captured by the written form of language, the emotional dimension is not – a question mark can direct a reader to read a sentence as a question, but there is no graphical correlate for the emotional dimension.

4.2 *Visualizing a Speaker's Voice*

4.2.1 *The Interplay Between Written and Spoken Language*

Although silent reading and spoken speech are at times thought of as different representations of the same thing, they are distinct in their perception and production [166], each with their own unique formation. While written language is related to spoken language, it is not a direct representation of it. Writing systems have to strike a pragmatic balance between information density and practicality. Take Hebrew script as an example. It uses diacritical marks to indicate vowels. These marks are often included in educational texts for children who are learning to read, but omitted in everyday writing, reflecting a trade-off between writing's practicality and the reader's ability to infer the missing information from context.

This inclusion or exclusion of elements within writing systems is tied to how people employ them. For instance, the presence of spaces between words, a seemingly essential feature in modern text, was only introduced in the 7th century A.D. because of how it helped guide the reading of Latin by individuals who did not know the language and, as such, needed an explicit demarcation between words [107]. This can sound counterintuitive, but, as says Seidenberg, 'there are no "spaces" between words in fluent speech' [166]. There are other examples of how mismatched written and spoken language can be: the phoneme, a seemingly natural unit when one considers written text, is not recognized in speech by those who do not know how to read [135].

While phonetic elements in the Latin alphabet are typically represented by letters, certain aspects of prosody can be conveyed through special marks. These include symbols that will direct emphasis or specific pitch patterns, *e.g.*, the question and exclamation marks. Others may denote rhythm, such as commas, periods, and dashes. Others, still, are more specialized, such as the *interrobang* (?!), the *love point* (♡), or the *doubt point* (⌘), but those are not widely used or available.

Importantly, the conventions that make these cues work are not static, and will change over time and/or intended contexts. McCulloch, for instance, has documented the many ways by which internet users have repurposed punctuation marks to convey tone and emotion in text-based communication – and that is even before we consider the use of emojis [127].

4.2.2 *Interventions to go from Expressive Textual Forms to Prosody-Rich Spoken Content*

Given that the written language has this malleability, many researchers have proposed different ways of augmenting text to overcome distances between the written and spoken forms of language. In some cases, these interventions are aimed at helping readers improve how they are able to convey in sound what they are reading.

In order to help disambiguate pronunciation of Dutch, for instance, Verbaenen created an expansion to the Times New Roman typeface that included various character variants that differentiated distinct phonemes that would otherwise be represented by the same letter [186]. In a similar vein, dos Reis developed the Speechant system, a prosodic annotation system overlaid on traditional text to help learners of English as a second language to better understand the rhythm and intonation of spoken English [58].

Bessemans et al. [22] have explored visual interventions on letter-shapes that convey elements of prosody, namely, pitch, rhythm, and loudness, to help novice readers improve their expressive reading skills.

All of these measures are aimed at fringe cases where a reader might find it challenging to read out aloud what they are reading. As such, they impose considerable additional effort on the writer, which is justified by the fact that these are not general purpose interventions.

4.2.3 *Interventions to Go from Prosodic-Rich Spoken Content to Expressive Textual Forms*

A separate line of work consists of capturing a specific ‘voiced’ reading and conveying it to a reader. As we discussed above, the same set of words can be interpreted in different ways depending on how they are spoken, and there are contexts where one might need to capture one such specific reading to convey it to a reader. Given the complexity of the processes involved, these approaches will usually involve some computational processing of an audio file that is able to extract expressive elements of speech, such as prosody, which can then be visualized. In the works discussed below, we focus on approaches that use typography to convey these expressive elements.

In some systems, authors have developed custom typographic-shaping systems that, rather than modulate existing typographic systems, build letter shapes computationally from the ground up based on prosodic features extracted from speech [158, 164, 205]. In others, they employ preexisting typefaces, but modulate traditional typographic parameters to echo prosody. Castro et al., for instance, mixed

discrete and continuous modulations by mapping loudness to font-weight, pitch to font-size, pauses as word-spacing, and word duration through letter repetition [38].

Other approaches aim to assist specialized audiences – *e.g.*, researchers in linguistics – 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 [6, 142].

In my own previous work, conducted during a prior master’s degree, I have explored hearing 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 [51]. This was followed up by a study that found that these modulations were enough to significantly differentiate between different readings of the same text [52]. While this work was not developed with DHH individuals, it served as a starting point to the studies presented in this dissertation.

4.2.4 *Going Beyond Prosody*

Some researchers have explored the depiction of features beyond those derived from the acoustic features modelled by prosody. Promphan, for instance, designed a typeface where the letter shapes were correlated to the valence of the words they represented, with negative valence words being depicted with spiky, harsh shapes, and positive valence words with soft, rounded shapes [149].

In a similar same vein, previous studies have also explored the use of color-emotion associations in written text to convey emotions [104, 170], such as employing red-colored text to express anger [177].

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 [69]. This system was later expanded to accommodate instant messaging scenarios [28].

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

These studies showcase varied approaches towards conveying prosodic and affective dimensions of speech through typography, but it is important to note that 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.

Indeed, while few studies have specifically explored prosodic or affective captions, there is a growing body of research in the fields of HCI and accessibility 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.

Some of these have explored changes to the visual form of captions, about which participants' preferences seem widely spread, hinting at there being room for experimentation, but that some users might be resistant to changes. Scene occlusion and legibility are critical considerations. In small-group online meetings, Berke et al. [19] saw a tension between caption styles that either favor legibility or less image occlusion, *e.g.*, captions with black boxes behind text versus those with transparent backgrounds. Crabb et al. [46]'s participants, having watched captioned videos inside a web page, had a slight inclination towards captions placed *outside* of the video area, giving high importance to avoiding image occlusion.

Other studies have investigated the importance of identifying the current speaker in a panel discussion [10], as well as the benefits of inserting correct punctuation or pauses in captioned videos to improve readability for DHH viewers [76, 193].

To improve the transparency of ASR processes, Berke et al. experimented with different visual strategies to represent the confidence levels that ASR systems assign to each word in a caption. They found that participants generally preferred unmarked captions and felt that the displaying confidence with a 'markup style was too distracting' [20]. Similarly, novel captioning styles, with or without iconic graphical overlays, were shown to help hearing speakers adapt their speaking behavior, potentially helping comprehension both of ASR systems and their DHH audiences [128, 168, 169].

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 [96]. 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 [18, 46]. Proper segmentation, which aligns caption boundaries with syntactic boundaries, has also been found to improve caption readability [193]. 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 [8, 9].

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, *i.e.*, captions that convey a speaker's emotions, have also been investigated, albeit tentatively. 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 [156].

In a recent CHI Late-Breaking work, Hassan et al. 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 [85].

* * *

As we have seen, paralinguistic cues aid in understanding speech. Because of this, and given how devoid of such cues captioned text renders speech, many researchers have looked at how to bridge the gap between spoken and written language through typographic innovations. These competing approaches do not present a clear case for whether DHH users would benefit more from captions depicting prosodic-based features or, alternatively, affective ones. As such, looking into this issue is the goal of Study 2,

presented in the next chapter. Study 3 follows with a systematic exploration of the design space of said paralinguistic features. Taken together, both studies aim to create a more nuanced and expressive captioning system that enhances comprehension and engagement for DHH viewers.

CHAPTER 5

Study 2: *What* Dimensions of Speech Should We Depict Through Captions?¹

This chapter aims to answer *what* dimensions of speech would be more helpful for DHH viewers if presented through visual cues added to captioned text. It contrasts a prosodic approach with an affective one, and in so doing, seeks to answer RQ2.A, namely, *How easily can a speaker's emotions, moods, and emphasis be identified when captions depict prosody and/or emotions in addition to words?* Additionally, we investigated whether the choice for these different approaches is informed by the setting where viewers use the enhanced caption styles. This was the focus of RQ2.B, namely, *In what settings is the use of visual depictions of prosody and/or emotion felt to be most appropriate from the point-of-view of DHH individuals?*

To answer RQ2.A, Study 2 exposed participants to three different types of captioning systems, each designed as the representation of a different set of prosodic or emotional features, thus gauging how each approach changed users' understanding of what was being said. A fourth type of caption, which had neither prosody nor emotions, was included as a baseline for comparisons. Additionally, to answer RQ2.B we measured participants' opinions about the ease of reading and appropriateness of each caption type for use in different settings.

5.1 *Study Design*

The test was online and self-administered. An introduction was done between one of the researchers and the participant over email and/or teleconferencing, after which the link to the test was shared. On

¹This study was part of a joint project between myself, Matthew Watkins, a graduate student at RIT, Dr. Sooyeon Lee, then a post-doc at RIT, and my co-advisors, Dr. Roshan L. Peiris and Dr. Matt Huenerfauth. I led the study design, stimuli creation, data analysis, and writing of the paper, which was published at the ACM CHI'23 conference [53].

this website, there was an introduction about the goals and workings of the study, with examples and explanations for the four types of captions (for details, see subsection 5.1.1 onward).

Following a demographic questionnaire, eight videos were presented on separate pages, with no sound, and captioned in one of the four available styles. While Python scripts pre-processed the speech files to extract affective and acoustic cues, a Javascript pipeline running on a web browser handled the styling and animating of the captions. HTML video provides a series of native events fired at key moments of each line of text's life-cycle (*cueEnter*, *cueExit*, etc), and it was through overloading these events that we were able to customize each caption style. Although we found that even mid-end machines, such as a 2014 Macbook Pro, were able to render the captions in real-time and virtually flicker-free, we felt it safer to present them 'burned-in' (*i.e.*, as open-captions) in the videos to account for participants' unpredictable computer settings.

The videos had an average duration of 50 seconds ($\sigma = 15$ s). As with Study 1, they were taken from SEND [143], and consisted of different individuals telling a personal story that had strong emotional overtones. To counter how each story could potentially bias participants' preferences, each video was generated in all four caption styles. Although all participants viewed the same eight videos, the order in which they were presented, as well as the caption style applied to each video, were randomized for every participant.

After watching each video, and immediately below it, participants indicated their agreement with the following statements on a 7-point Likert scale: (1) I found the speaker's emotions and moods easy to identify; (2) I could easily tell which words were emphasized; (3) I would be interested in using this captioning style for *work meetings* in software such as Zoom, Google Meet, etc; (4) I would be interested in using this captioning style for *personal meetings*; (5) I found the captions easy to read.

After watching the eight videos, participants were asked a set of open-ended questions which were chosen according to *that participant's* specific answers to the Likert-scale items shown previously. The goal here was to add nuance to our understanding of the quantitative answers collected. For instance, if for a given video the participant gave a below mid-point rating to the scale '*I found the speaker's emotions and moods easy to identify*,' an image of that same video would resurface at this last step with the prompt: '*Could you elaborate on why you felt that the caption design shown was not helpful to understand the speaker's emotions and moods?*'

well, we had been having some relationship problems and we were just both really stressed with college applications and what not,

(a) Conventional style (C)

well, we had been having some relationship problems and we were just both really stressed with college applications and what not,

(b) Prosody style (P)

well, we had been having some relationship problems and we were just both really stressed with college applications and what not,

(c) Emotion style (E)

well, we had been having some relationship problems and we were just both really stressed with college applications and what not,

(d) Prosody & emotion style (P+E)

Figure 5.1: The four caption styles used in the test. The C style has words with no additional styling (5.1a); The P style maps loudness to font-weight, pitch to baseline shift, and duration to letter-spacing (5.1b); The E style maps valence to color, with red meaning negative, white neutral, and green (not shown) positive, and arousal to font-size (5.1c); finally, the P+E style combines the five modulations from both the P and E styles (5.1d).

5.1.1 *Extracting Prosodic and Emotional Features*

The four different sets of features represented were: (1) No prosodic or emotional features – basically, a conventional captioning system (identified in this chapter as style C) depicting only words in a neutral fashion – shown in Figure 5.1a; (2) Only prosody (style P), shown in Figure 5.2; (3) Only emotions (style E), shown in Figure 5.1c; and (4) a combination of Prosody and Emotions (style P+E), shown in Figure 5.1d. We describe how the data depicted in styles P (2), E (3), and P+E (4) was acquired below, with the visual design for the representations discussed in subsection 5.1.2.

5.1.1.1 *Force-Alignment*

The SEND dataset includes not only the set of videos previously described, but also a transcription of all spoken content, grouped in 5-second blocks of words. For our purposes, this granularity was insufficient, given how we needed timestamps of each individual word to extract its prosody and emotions. To obtain these, we fed the transcriptions into an instance of Gentle [141], a Kaldi-based force-alignment toolkit set at a word-based granularity level.

5.1.1.2 *Extracting Prosody from Speech*

We followed the extraction and processing procedures I previously developed with Dr. Paula Dornhofer Paro Costa, as outlined in de Lacerda Pataca and Costa [52]. Loudness and pitch were extracted in the Praat software [30], patched through Python using the praatIO interface [121]. The algorithms used were root mean square for loudness and auto-correlation for pitch, applied to each segmented words' audio file. The duration of this file also gave us the raw-duration values for each word.

As in de Lacerda Pataca and Costa [52], we normalized the raw values of loudness and pitch using both local and global ranges. The local range considered 10 words before and 5 after the target word, while the global range included all words in the video. We then averaged the normalized values calculated from these ranges. This asymmetric local window reflects how prosodic features are perceived contrastively within short stretches of surrounding context rather than in isolation [14, 47]. Visual typographic changes, however, are somewhat different: the visual scale has clearer absolute bounds (*e.g.*, how thin or thick a font can become). For this reason, we balance the local with a global range of values, as I detailed in de Lacerda Pataca [49]. With $x_{n \text{ raw}}$ as the n th raw measurement of feature x , the normalized value $x_{n \text{ norm}}$ is calculated as:

$$x_{n \text{ norm}} = \frac{1}{2} \left(\frac{x_{n \text{ raw}} - x_{\text{local min}}}{x_{\text{local max}} - x_{\text{local min}}} + \frac{x_{n \text{ raw}} - x_{\text{global min}}}{x_{\text{global max}} - x_{\text{global min}}} \right) \quad (5.1)$$

For rhythm, however, we had to modify de Lacerda Pataca and Costa [52]'s syllable-length-based algorithm, given that it was originally applied to Brazilian Portuguese and our experiment would be in U.S. English. Brazilian Portuguese is considered a syllable-timed language in certain conditions [130], *i.e.*, one can expect the duration of each syllable to be roughly similar, which gives the language a machine-

gun-like rhythm. When there *are* changes in this average duration, one can assume a meaningful change in prosodic rhythm. U.S. English, on the other hand, is stress-timed, *i.e.*, the regularity is in the rhythm between stressed syllables, giving the language a morse-code-like rhythm [139]. In the case of U.S. English, syllables will naturally have different durations, so the same syllable-duration metric would not be as effective as a marker for prosodic rhythm.

To work around this, we used a text-to-speech synthesizer² to sound each word in the transcription. This synthetic word's duration was then used as a normalizing denominator for the duration of the actual spoken word, *i.e.*, how faster or slower the actual spoken word was when compared to this 'neutral' (but consistent) synthetic counterpart. Because the same word always has the same duration when produced by the synthesizer, this approach isolated fluctuations in speed in the spoken signal. This new metric was thus used as a correlate of prosodic rhythm.

5.1.1.3 *Extracting Emotions from Speech*

Building on the work of Russell [160], we modeled emotions using two dimensions: valence and arousal. Valence represents the pleasure-displeasure axis, while arousal reflects the level of activation or alertness. These dimensions are typically visualized as, respectively, x, y coordinates on a two-dimensional plane, where different positions correspond to distinct emotional states. For instance, excitement is characterized by high valence and high arousal, whereas sadness is associated with low valence and low arousal.

To obtain valence and arousal values, we ran the segmented audio recordings through a transformer-based neural network [190], which outputted values for valence and arousal. Notwithstanding its state-of-the-art accuracy, we chose this particular model because of two key characteristics: (1) it can generalize better across domains, *i.e.*, even if not trained to our specific dataset³ it is expected to lose less accuracy than alternative architectures; and (2) its accuracy suffers little loss between speakers of different genders, as were present in the dataset used.

The model was trained on sequences between 3 to 10 seconds long, so it would not be able to extract meaningful values from the audio sliced in too-short word-sized chunks. Following a suggestion from

²Using the Python library `pyttsx3` [23], we created an instance of macOS' native speech synthesizer, `NSSpeechSynthesizer`, set with the default voice.

³Interestingly, while the only inputs used were the audio files, it has been shown that this particular architecture is implicitly able to derive affective information from linguistic elements present in speech, helping it beat the performance of explicitly multimodal neural networks [185].

one of the authors (as communicated via email [184]), we padded each word with its surrounding audio, adding a 3-second margin on each side.

5.1.2 *Caption Design*

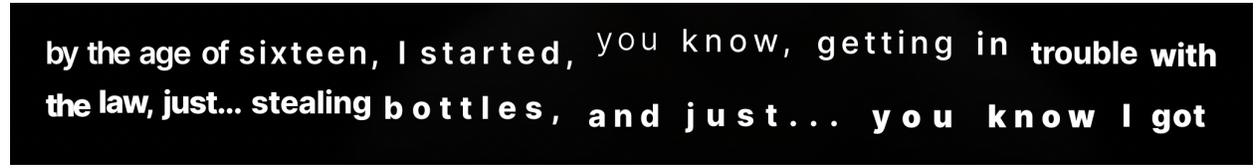
There were two main constraints to the visual choices we explored when designing the speech-modulated typography for the captions. First, we focused on purely typographic designs, which, as we saw in Study 1, were less distracting than independent interface elements. While this is not an empirically validated point, we use it here as a simplifying assumption to narrow our focus into a more manageable, albeit still vast, subset of visual approaches.

Second, we limited ourselves to typographic parameters that are freely combinable. Because we had a caption style where *both* prosody and emotions are shown at the same time, visual solutions applied individually to prosody and emotion must allow for a third, combined setting where all five different input dimensions (three for prosody and two for emotions) vary simultaneously.

5.1.2.1 *Depicting Prosody: Loudness, Pitch, and Duration*

Unlike with emotions, there are quite a few competing models that represent prosody through typography [22, 38, 51, 52, 158, 164, 205]. We followed an approach I developed previously [52] because of its flexibility in allowing the modulation of extra typographic parameters, as needed for depicting emotion.

Figure 5.2 shows an example of prosodic-based captions, where we mapped loudness to font-weight (thin to thick equating quiet to loud words), pitch to baseline shift (negative to positive vertical displacements equating low to high-pitched words), and duration to letter-spacing (tight to spread out letters equating fast to slow sounding words).



by the age of sixteen, I started, you know, getting in trouble with the law, just... stealing bottles, and just... you know I got

Figure 5.2: Example of the three prosodic features mapped as modulations of three typographic parameters applied to a block of captions.

5.1.2.2 *Depicting Emotions: Valence and Arousal*

Challenges related to the ambiguity of captioned speech were a leading theme in Study 1. For Study 2, however, our design goal was not to create a model of captions that would remove all ambiguity in speech but, instead, one that would provide users with tools to better deal with it. Thus, while previous authors have worked with categorical models that define explicit emotions [88, 112, 155], our premise led us to work with a dimensional model.

While categorical emotions can map to the circumplex plane, these mappings may denote distinct ‘categorical’ emotions depending on the context (*e.g.*, fear and anger are both emotions with negative valence and high arousal) [157]. In fact, in Russell’s original paper many similar models are shown to have existed, all giving slightly different, albeit related, meanings to the two axes [160].

A design goal of embracing ambiguity is based on the assertion that to understand emotions in speech one must consider that meaning is not *only* found in an acoustic signal but also in how this signal is grounded in a particular socio-cultural context [29]. Leaving room for ambiguity can thus be thought of as an asset, *i.e.*, a recognition that one’s interpretation of something can vary depending on the subjects involved and the context [72]. By embracing how the ambiguous nature of emotions gives form to open-ended visual representations, we align ourselves with Höök et al.’s design principle of non-reductionism, and Boehner et al.’s call for systems that support interpretive flexibility [29, 87].

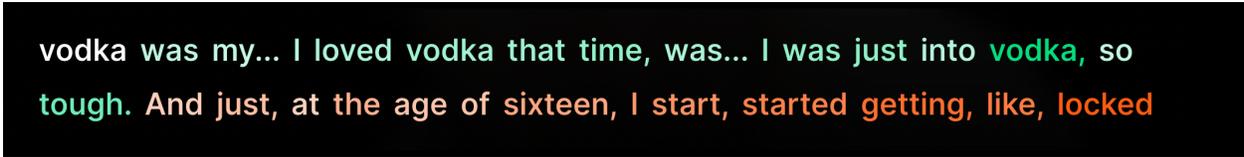
All of this, however, does not equate to leaving individuals confused. As such, our design choices aim to be intuitive representations of both valence and arousal. The literature points to two common approaches: either using animation effects, especially if mimicking the bodily expressions of emotions [69, 122, 156], or directly manipulating type shapes, be it programmatically [118] or directly in their typographic designs [149].

Word animations were deemed unsuitable; they tend to fragment lines of text, which is problematic given that automatic speech recognition systems use scrolling captions where word positions continuously shift with the appearance of new lines. This creates an unpredictable compounding motion when combined with ‘affective’ animations. Conversely, the method of embedding valence in the type shape, particularly Promphan’s *Emotional type*⁴ [149], appears more suitable for captions. However, it does

⁴The author kindly provided us with a revised version of the typeface published in her thesis, which served as the basis for one of the prototypes presented to participants in Study 1 (see Figure 2.2c).

not meet our second design constraint for this study: the need for typographic parameters to be freely combinable. Thus, neither approach could be utilized effectively.

The chosen typographic parameters were color, for valence, and font-size, for arousal. We based our choice for color on ample evidence of how it can be used to represent moods and/or emotions [15, 40, 41, 73, 104, 112, 174]. While the use of red to represent negative valence seems a relatively straightforward choice [73, 104], we saw evidence for the use both of blue [174] and green [104] to represent positive valence. In Study 1 we pilot-tested a red-to-white-to-blue scale color scale to represent, respectively, negative to neutral to positive valence, but negative feedback from participants led us to settle on a red-to-white-to-green scale for the second study. While this color scheme is hard to distinguish for individuals with severe types of red-green color vision deficiency (protanopia and deuteranopia), we tinged the red with some yellow and the green with some blue to make them more discernible to individuals with the more common, milder deuteranomaly and protanomaly [138]. Figure 5.4 shows a simulation of how the palette weathers under different types of color vision deficiency⁵, and Figure 5.3 shows an example of this color scheme applied to captions.



vodka was my... I loved vodka that time, was... I was just into vodka, so tough. And just, at the age of sixteen, I start, started getting, like, locked

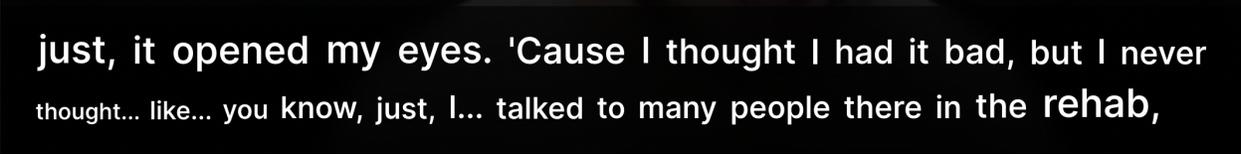
Figure 5.3: Font-color representing valence.



Figure 5.4: The valence color palette under simulation of various types of color vision deficiency.

To the best of our knowledge, there are no direct examples of the representation of arousal by the modulation of typographic parameters. We opted to use font-size because it has seen use as a representation for both changes in pitch [38] and loudness [158], features which have been associated with emotions of high (joy, anger) or low (sadness) arousal [57, 102]. An example of this modulation can be seen in Figure 5.5.

⁵Images created in the Coblis Color Blindness Simulator [200]



just, it opened my eyes. 'Cause I thought I had it bad, but I never thought... like... you know, just, I... talked to many people there in the rehab,

Figure 5.5: Font-size being used to represent arousal.

5.2 *Participants and Recruitment*

The IRB-approved study ran from July to August 2022. \$40 compensation was offered. Recruitment was done through DHH specific social-media channels and mailing lists, screened for by identification as a DHH individual and not having participated in Study 1. A total of 16 individuals took the test. Eight identified as male and eight as female. Nine identified as being deaf/Deaf, and seven as hard-of-hearing. Their average age was 26 ($\sigma = 5$). Asked about how comfortable they were with reading and writing English, participants' median answer (on a Likert scale going from 1 to 7) was 7. Participants completed the test on average in 32 minutes ($\sigma = 14'$).

5.3 *Findings*

We conducted statistical significance testing on the Likert responses using a Kruskal-Wallis test, which was significant for all distributions. We then ran a post hoc Mann-Whitney U test between each caption type, with p-values adjusted using Holm-Šídák corrections. Figure 5.6 shows the distribution of answers, which comparisons were significant, and the median score for each scale.

5.3.1 *Quantitative Data*

Median responses for the *clarity of emotions and moods* Likert scale for the C, P, P+E, and E styles were, respectively, 4, 4.5, 6, and 6, with statistically significant differences between the C and E styles ($UT = 295.0$, $p < 0.05^6$), P and E styles ($UT = 252.0$, $p < 0.01$), and P+E and P styles ($UT = 734.0$, $p < 0.05$).

⁶P-values presented adjusted using Holm-Šídák corrections. Always two-tailed tests with n_1 and n_2 equal to 32.



(a) Key to the responses to the Likert-type scales.



(b) Answers to the Likert scale presented immediately below each video with the following statement: *I found the speaker's emotions and moods easy to identify.*



(c) As above, but for: *I could easily tell which words were emphasized.*



(d) As above, but for: *I would be interested in using this captioning style for work meetings in software such as Zoom, Google Meet, etc.*



(e) As above, but for: *I would be interested in using this captioning style for personal meetings in software such as Zoom, Google Meet, etc.*



(f) As above, but for: *I found the captions easy to read.*

Figure 5.6: Responses to the five Likert-type scales. Each row represents responses considering one out of the four caption styles. Caption styles are abbreviated as follows: Conventional (C), Only prosody (P), Prosody and emotions (P+E), and Only emotions (E). A blue ** marks a $p < 0.01$ significant Mann-Whitney U test between medians, while a black * marks a $p < 0.05$ comparison.

Median responses for the *clarity of emphasis* Likert scale for the C, P, P+E, and E styles were, respectively, 3, 5, 5.5, and 5, with statistically significant differences between the C and P+E styles ($UT = 302.5, p < 0.05$), and C and E styles ($UT = 269.0, p < 0.01$).

Median responses for the *use in work meetings* Likert scale for the C, P, P+E, and E styles were, respectively, 6, 3, 3.5, and 5, with statistically significant differences between the C and P styles ($UT = 801.0, p < 0.01$), C and P+E styles ($UT = 748.0, p < 0.01$), and P and E styles ($UT = 287.0, p < 0.01$).

Median responses for the *use in personal meetings* Likert scale for the C, P, P+E, and E styles were, respectively, 6, 3, 4.5, and 5, with statistically significant differences between the C and P styles ($UT = 790.5, p < 0.01$), C and P+E styles ($UT = 735.5, p < 0.05$), and P and E styles ($UT = 291.5, p < 0.05$).

Median responses for the *legibility* Likert scale for the C, P, P+E, and E styles were, respectively, 7, 4.5, 5, and 6, with statistically significant differences between the C and P styles ($UT = 838.0, p < 0.01$), C and P+E styles ($UT = 806.5, p < 0.01$), C and E styles ($UT = 740.0, p < 0.01$), and P and E styles ($UT = 311.5, p < 0.05$).

In summary, captions that had an emotional component (the P+E and E styles) significantly outperformed conventional captions in how they were able to help participants identify emphasis, but only the E style showed a significant improvement on how it made *emotions* and *moods* easier to identify. Traditional captions were perceived as more legible than all three other styles, including E, which outperformed P. Participants were less interested in using either the P+E or P styles than traditional captions for workplace or personal meetings – for E captions, the preference is smaller, but is not significant.

5.3.2 *Open-Ended Comments*

After watching the videos, participants were asked questions about what worked or did not in the caption styles, with specific questions being chosen according to their most negative and positive answers to the Likert-type scales.

Regarding legibility, there were comments about how the new styles (P, P+E, and E) were harder to read than the more traditional C style. Prosodic representation, in particular, was disliked, with specific mention of how its use of baseline shift and changes in the spaces between letters made words ‘wild’ and hard to read. Some participants also had the impression that there was sometimes too much going on, making captions confusing, slower to read, or even headache-inducing.

Speaking specifically to the typographic parameters modulated in the three new styles, some participants commented on how font-weight and color worked effectively to represent a speaker's emotions and moods. Changes to font-size were also positively cited as expressive modulations, with the caveat that at times they made captions too small to read comfortably. Lastly, and more rarely, a few participants felt changes in baseline shift and letter spacing negatively impacted legibility.

In terms of function, and particularly regarding styles P+E and E, the new captions received praise. They were said to work in terms of making the speakers' emotions and speech clearer. One participant imagined this reducing misunderstanding their friends. Another said that, while they are typically able to derive sufficient emotional understanding from facial expressions, the E-styled captions would be 'awesome' when the speaker's face is occluded. Discussing style E, one participant said it was easier to tell when the mood shifted between positive or negative feelings, which another participant said changed how they understood the stories in the videos.

More broadly, some comments pointed out that, even if the E style did not necessarily change their understanding, it was less bland than traditional captions. Similarly, P+E was felt as being more engaging and easier to follow along. Some participants claimed that while they personally did not see a need for these new styles, they felt other DHH individuals could benefit from them.

5.4 *Discussion*

Results revealed that the Emotion (E) caption style significantly outperformed conventional captions in participants' perceptions of how they expressed emotions and emphasis, with the Prosody + Emotion (P+E) variant also scoring higher in its depiction of emphasized words. This is an encouraging suggestion that the E style could help make accessible these important paralinguistic dimensions of speech.

Surprisingly, the Prosodic (P) style did not outperform traditional captions in representing emphasis. Given that it was based on a model that had been shown to successfully depict speech prosody with hearing individuals, it raises the question of how differently hearing and DHH individuals will interpret these captions.

We saw that both depictions with prosodic components had relatively low legibility scores, with two of its three typographic parameters (baseline shift and letter-space) being specifically denounced by some participants as culprits. Given that it also had low interest in use for personal or work meetings, the

E style's performance for emotion and emphasis representation, coupled with its higher legibility and appeal, gives an unexpected but interesting answer to our RQ2.A and RQ2.B:

In seeking what dimensions of paralinguistic properties could represent a speaker's emotions and/or emphasis in captioned speech (RQ2.A), it seemed plausible to expect that the P style would score higher at representing emphasis and the E style at emotions. We found, however, that a choice between showing either emotions or emphasis, as came out of Study 1, may be unnecessary, with the E style capable of capturing and representing both dimensions.

As for RQ2.B, while the E style did not outperform traditional captions in perceived suitability for work or personal meetings, it ranked significantly better than the P style for both settings. The assumption that there would be a divide between what was favored for work versus personal meetings did not pan out, with each style's preference score relatively consistent between the two settings (this effect could, however, be an artifact of how the videos we used skewed towards content one would expect to see in a personal conversation instead of a professional one).

5.5 *Limitations and Future Work*

Study 1 (Chapter 2) showed that further work was needed both towards a better understanding of *what* dimensions of speech should be visualized and *how* to design those visualizations. With Study 2, we investigated the first path, but both are intertwined, *i.e.*, to test *what* to represent, we needed to design the options somehow, and inversely, if we were to evaluate different caption designs, they would have to depict some model of these dimensions. As such, while we based our design choices on fair assumptions from research related to ours, given that the field is still sparse, further work is needed to investigate *how* to represent these dimensions systematically. In fact, this question was a central theme in Study 3, presented in Chapter 6.

This design process needs to consider the perspective of DHH individuals, taking into account two considerations that emerged from Study 1 and 2: First, the color schemes used may leave individuals with color vision deficiency unable to distinguish between negative and positive valence words, as is seen in the last two rows of Figure 5.4. Future research should explore alternative color schemes and typographic modulations to make the style more accessible.

Second, legibility was a recurrent concern for participants from both Study 1 and 2. We used a Likert-rating scale to weed out acute issues with ease of reading with any of the three proposed caption styles. This was, however, a somewhat blunt instrument, ignoring aspects such as gaze time (which is already high for DHH individuals for conventional captions [94]), reading speed, cognitive workload, etc. A caption design's readability is related to users' demographics, personal preferences, and use cases, so a *one-size fits all* solution is probably not an ideal approach here [16]. Still, as these designs mature, further research should investigate their reading performance more thoroughly. This may include exploration of different granularity levels for the measurement and display of the non-textual dimensions of speech – while, like Rosenberger-Shankar and MacNeil [158], we employed a word-level measure, this could be finer, *e.g.*, syllabic [52, 158] or phoneme-grapheme mapping [205], or coarser.

From speaker's perspective, having an autonomous system that *proactively* codifies and depicts their speech based on an automatic analysis of their emotions carries the risk of a loss of autonomy, as described in Höök et al. [87]. Future studies should investigate how a user interface could represent and, if needed, cede control to speakers about this emotion-sensing process, as it is ongoing.

We asked participants how clearly they could perceive emotions, moods, and emphasis in a captioned video. We did not measure, however, how helpful these represented dimensions were. Future studies could investigate whether the presence of these novel caption styles could alleviate ambiguity, especially considering the communication breakdown scenarios our interviewees described in Study 1: quick shifts in mood, inexpressive body language, and occluded faces, among others.

This last factor, of how a clear view of the speaker's face can influence how these captions are understood, might be an independent line of research of its own, *e.g.*, [7], given that, in Study 2, videos showed speakers' faces clearly, a condition which plausibly could affect the interpretation of the captions themselves (*e.g.*, similarly to how individuals with cochlear implants derive greater benefit from synchronous facial and speech channels than hearing individuals in emotion recognition tasks [189]).

While it was not a measured dimension, the fact that some participants of Study 2 commented about how the new captioning styles were more engaging is a compelling counterpoint to how some interviewees in Study 1 complained that traditional captions are *boring*. Considering that one of the comments specifically said that the new captions did not change their understanding of the story but were easier to follow along, we can envision that future studies could focus on measuring how immersive these captions are compared to past measures of traditional captions [103]. Indeed, this is an important focus of our proposed Study 4, presented in Chapter 10.

5.6 *Conclusion*

In this study, we contrasted how three novel caption styles represented emotions and emphasis compared to traditional captions. We found that the best-performing option was based on the output of speech processed through a neural network that extracted emotional features from it. This approach had good legibility, albeit worse than conventional captions. Participants' willingness to use these captions for work or personal meetings is comparable to that of traditional captions.

Our work has investigated a rarely explored dimension of DHH experience with automatic captions, putting forward three novel approaches for modeling, processing, and depicting speech that may motivate the development of more inclusive captioning systems.

CHAPTER 6

Study 3: *How to Depict Emotions in Affective Captions?*¹

6.1 *Introduction*

We saw in Study 1 that the lack of paralinguistic cues in captions can make captioned speech ambiguous and hard to parse. Study 2 investigated which speech features are most important to convey in captions, with an edge found towards depicting emotions, *i.e.*, a speaker's valence and arousal levels.

The challenge, then, becomes *how* to design these 'affective' captions. As we have seen, many researchers have explored conveying paralinguistic information through typography [22, 51, 52, 170, 205], but this prior work offers little guidance here because of how it has mostly focused on conveying prosodic aspects of speech, such as its pitch, rhythm, or loudness. What little research there is (*e.g.*, [99]) does not focus primarily on DHH individuals' perspectives on what caption styles are preferred and perform better at conveying emotions.

In this chapter, we address this gap by reporting on Study 3, a three-phase study 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:

¹This study was part of a joint project between myself, Dr. Saad Hassan, an assistant professor at Tulane University, Nathan Tinker, an undergraduate student at RIT, and my co-advisors, Dr. Roshan L. Peiris and Dr. Matt Huenerfauth. I led the study design, stimuli creation, data collection and analysis, and writing of the paper, which was published at the ACM CHI'24 conference [55].

- In Phase 1 (Section 6.2), 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 Phase 2 (Section 6.3), we examined six combinations of styles (based on the winning styles from the first phase) 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 Phase 3 (Section 6.4), we compared the four top-performing styles from the previous phase 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 phases 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 6.1 illustrates this process.

Ultimately, we sought to answer the following research questions:

- RQ3.A Are there caption styles that emerge as preferred by DHH viewers to represent valence or arousal *when depicted individually?*
- RQ3.B Are there caption styles that emerge as preferred by DHH viewers to represent valence or arousal *when depicted in combination?*
- RQ3.C What factors influence DHH viewers' preference for specific caption text styles conveying valence and arousal in speech?
- RQ3.D Do the most preferred methods for conveying valence and arousal in combination, selected in the answering of RQ3.B, outperform a baseline caption text when DHH participants *engage in an emotion-recognition task when watching captioned videos?*

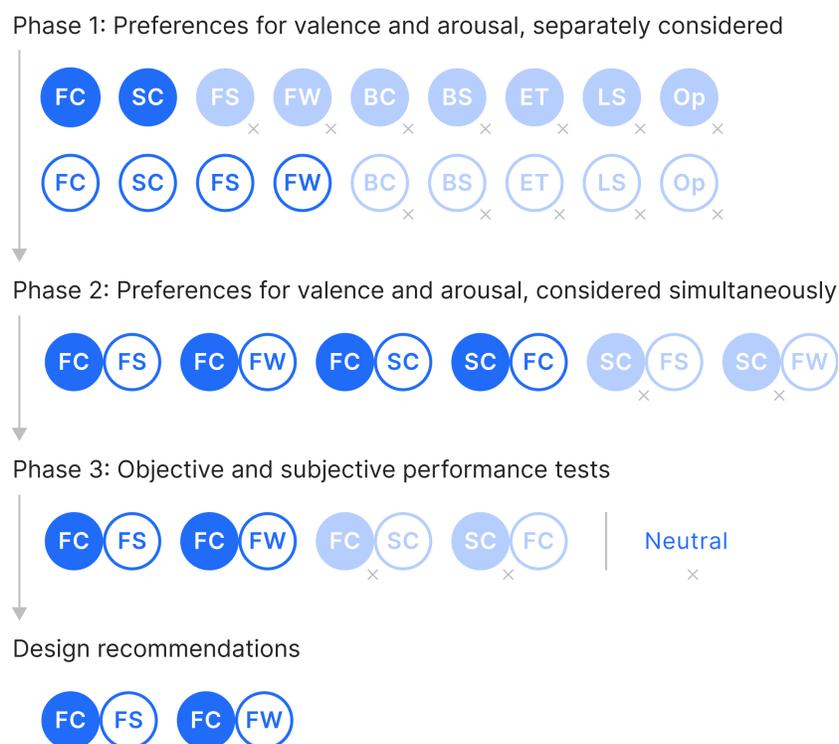


Figure 6.1: Map of the three phases and our design recommendations. Valence styles are represented by filled circles and arousal by outlined circles. The typographic modulations are abbreviated as follows: BC: background-color; BS: baseline shift; ET: emotional typeface. FC: font-color; FS: font-size; FW: font-weight; LS: letter spacing; Op: opacity; and SC: shadow-color; Each phase eliminated certain styles from consideration, marked by faded colors and a cross symbol (×).

RQ3.E Do the most preferred methods for conveying valence and arousal in combination, selected in the answering of RQ3.B, outperform a baseline caption text when DHH participants *report on their subjective impressions of how each caption style performs according to the factors outlined in the answering of RQ3.C?*

6.2 Phase 1: Evaluating Captions Styles that Depict Valence or Arousal Individually

In the first phase of Study 3, 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 6.3). 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 (RQ3.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 (RQ4).

6.2.1 *Methods*

6.2.1.1 *Stimuli Generation*

6.2.1.1.1 *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. As in Study 2, we used videos from the Stanford Emotional Narratives Dataset (SEND) [143]. 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 [141], 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 [190] to obtain values of valence and arousal, emotional components as defined by the circumplex model of emotion [160].² These two values, which are the final output of this process, as illustrated in Figure 6.2, were then normalized [52] and annotated in a caption file [49].

6.2.1.1.2 *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 emo-

²We acknowledge the scholarly debate challenging the possibility of there actually being a meaningful ground *truth* for such a system to deduce [29, 87], 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.

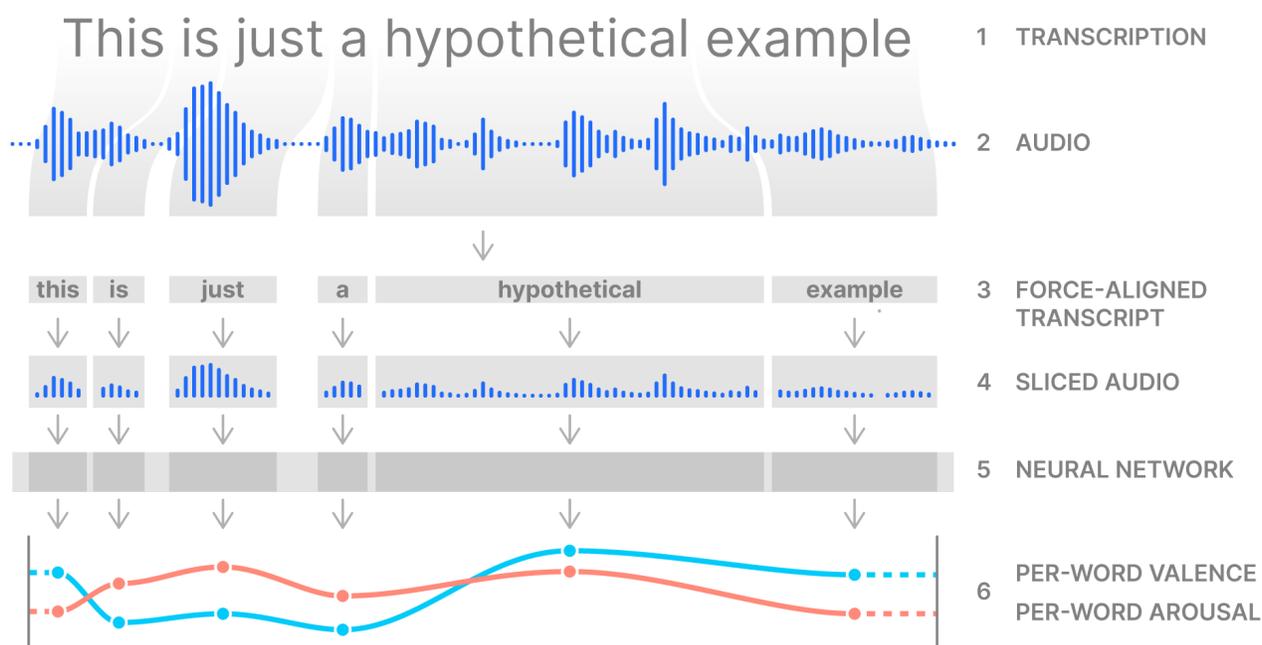


Figure 6.2: Diagram of how the transcription of a spoken utterance (1), together with its audio file (2), are used to generate a force-aligned transcript (3), which allows for the slicing of each word's audio (4), which is then fed into a neural network (5) that outputs its valence and arousal levels (6).

tional 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 [45] 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 [15, 40, 41, 53, 73, 85, 104, 112, 174]. Frequently, this is used to depict valence, with red commonly associated with negative values [73, 104] and blue [174] or green [104] used for positive ones. We used the color palette defined in Hassan et al. [85], designed to ensure that individuals with more common forms of 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 6.3a) style involves changing the color of the word itself [53]. *Background-color* (Figure 6.3b) involves adding a colored

box behind each word. This is similar to visual experiments done for instant messaging interfaces [40, 41]. Finally, the *shadow-color* (Figure 6.3c) 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 6.3d, 6.3e, and 6.3f, respectively) have been used by various authors to represent elements of prosody, such as loudness, pitch, and duration [22, 38, 51, 52, 158, 164, 205], or arousal and intensity of valence [85].

Font-size (Figure 6.3g) has been used to depict arousal [53]. Hassan et al. [85] used brightness to depict dominance, which we here adapted as *opacity* (Figure 6.3h). Lastly, Promphan [149]'s *emotional-typeface* (Figure 6.3i) 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 [11].

6.2.1.1.3 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.



Figure 6.3: The nine caption styles used in the first evaluation. All images are screenshots from one of the videos that were used as stimuli.

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.

6.2.1.2 *Evaluation Design and Analysis Plan*

Answering RQ3.A 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 were whether each style did a good or bad job at depicting either valence or arousal. This process is shown in Figure 6.4. 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 differences³ between items [21]. 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 [21] or, more broadly, that measure typographic style preferences [192]. Lastly, BWS is more robust than Likert-rating scales against inconsistencies⁴ in participants' ratings across multiple rounds [48, 100].

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

⁴Repeated 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.

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.

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 [61]. 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 implementation used was Herbrich et al. [86]'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 . It quantifies both a style's expected performance and the level of uncertainty around this estimate, allowing us to moderate the degree of trust in the results. Larger or smaller values of σ reflect more or less uncertainty in the overall rankings, which is expected to diminish as more matches are run and if participants' answers are convergent. To determine a given caption style's true relative strength, we use $\mu \pm 2\sigma$, providing a 95% confidence interval.⁵

6.2.1.3 *Experimental Set-up*

Both the first and second phases of Study 3 shared a similar overall structure. In Phase 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 ($\sigma = 4s$). Figure 6.5 shows a screenshot of the interface used during the first phase of the study. As with the following two phases, this experiment was developed as a website using jsPsych [56].

⁵A 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. [42]'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.

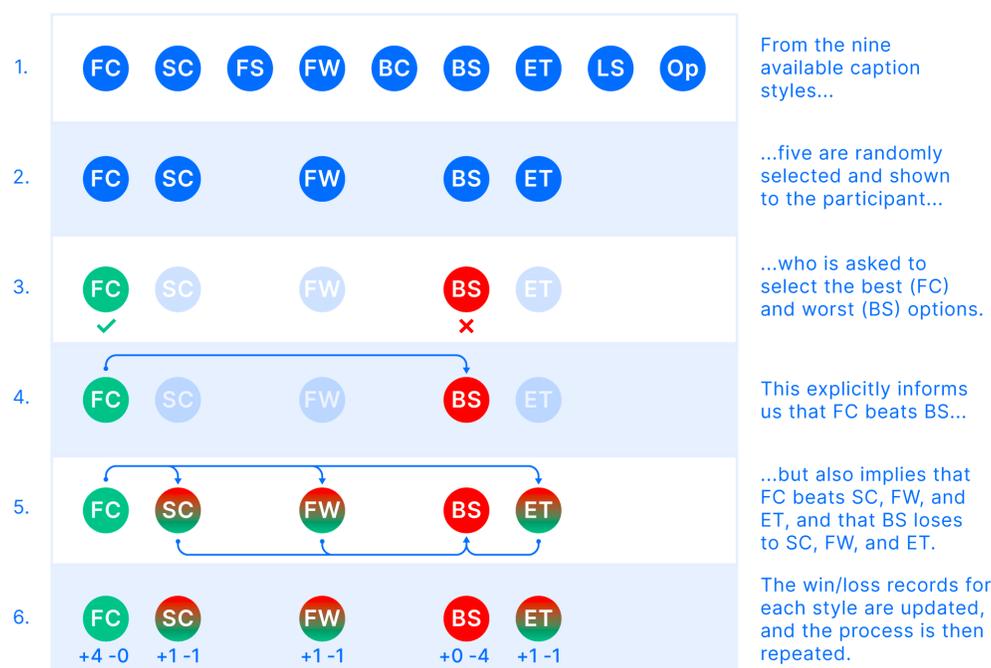


Figure 6.4: Example of one round in our Best-worst scaling setup. Caption style names are abbreviated as in Figure 6.1.

In Phase 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?*

6.2.2 Findings from Phase 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 Deaf or Hard-of-Hearing. For Phase 1 we recruited a total of 10 participants, 7 of which identified as female and 3 as male, 7 of which identified as Deaf and 3 as Hard-of-Hearing, with a mean age of 29.5 years ($\sigma = 11.9$).

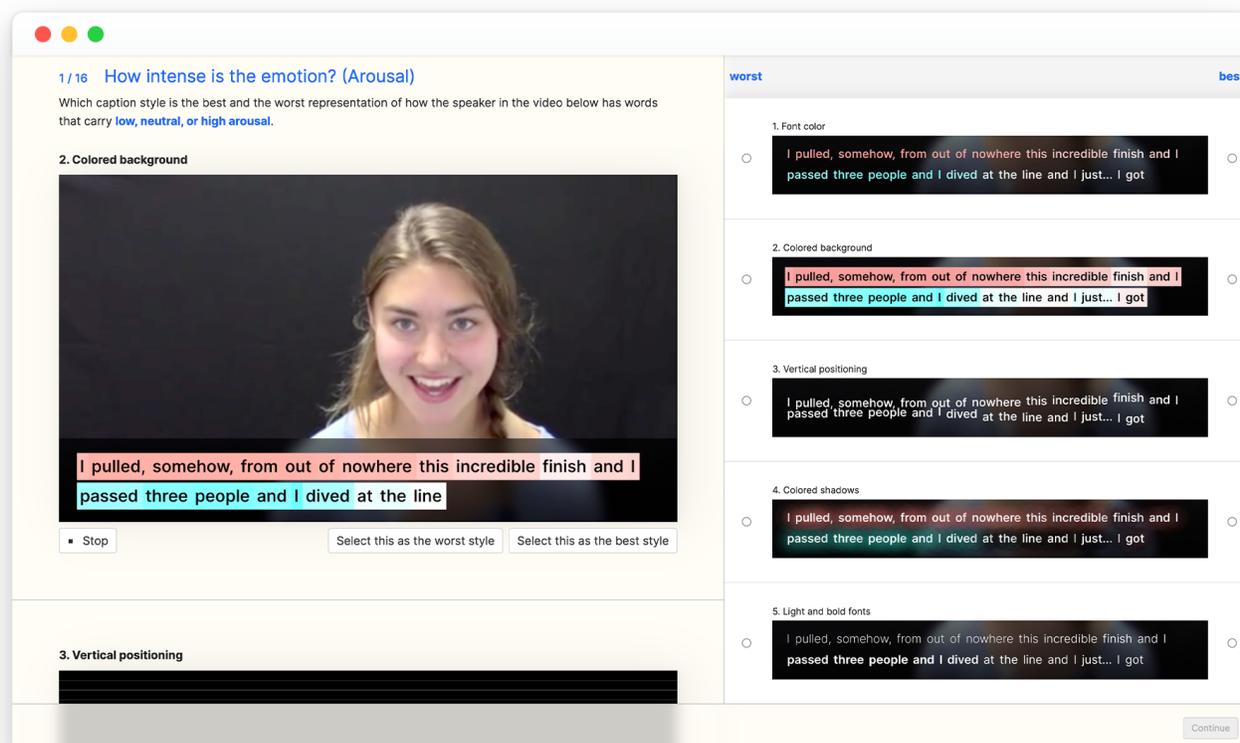


Figure 6.5: Screenshot of the experiment's platform. On the left side of the image, an example video is shown with the background-color caption style. On the right side, five different caption styles are displayed, which were presented to participants in a particular Best-worst scaling (bws) round. The 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.'

6.2.2.1 Caption Style Reference Rankings

Phase 1 had 10 participants evaluating 5 videos per round for 8 rounds for each affective dimension. A five-way BWS generates 7 data pairs, so $10 \times 8 \times 7 = 560$ pairwise comparisons for both valence and arousal. Table 6.1 shows the results from this first phase of 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.

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

Table 6.1: 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.

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 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 [101]. The data, transformed into scores using the counting procedure outlined by Orme [144], 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:

⁶See Lee [113] 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 $\tau = \sigma/100$.

- 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 ($\mu = 21.5, \sigma = 0.9$).

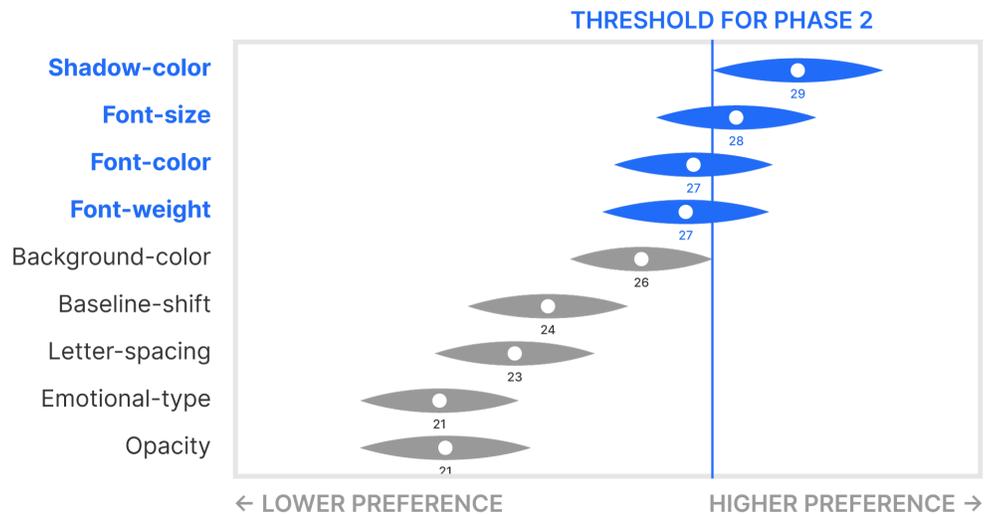
Styles in bold were included for Phase 2 if their higher bound was greater than the lower bound of the top-scoring style. TrueSkill results are also shown in figure 6.6.

6.2.2.2 Open-Ended Feedback from Participants

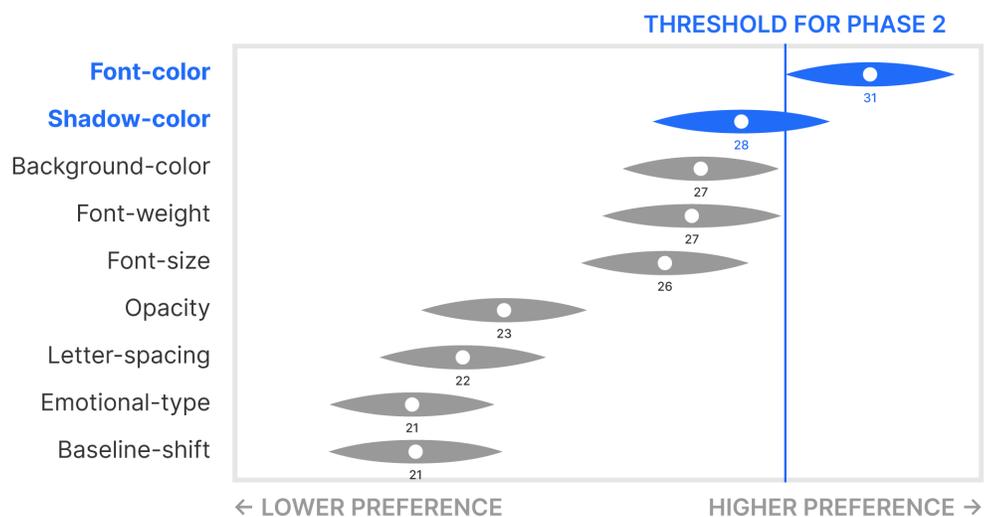
After the BWS part of the test, participants were shown, for each of valence and arousal, the caption styles that best and worst performed according to their individual votes. They were then asked to comment on the reasons they thought these styles were (or were not) able to convey valence or arousal. (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. 6.3i), P1 said it did not 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] did not 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. 6.3e) did not 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.'*



(a) Preferences for valence-representing caption styles



(b) Preferences for arousal-representing caption styles

Figure 6.6: Charts showing the relative strength and confidence interval for each caption style in relation to valence and arousal, using data collected in Phase 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 Phase 2.

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. 6.3h) 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. 6.3g) 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. 6.3a) 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. 6.3b) 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 were harder to recognize.*

Participants felt shadow-color (Fig. 6.3c) 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.*

6.3 *Phase 2: Evaluating Caption Styles that Depict Valence and Arousal in Combination*

After completing Phase 1, we used its results to identify a small number of combinations between the most preferred caption styles for valence and arousal, which we would now evaluate in a new study session with new participants. 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 mappings between affect and typographic 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 Phase 2, as seen in Figure 6.7.

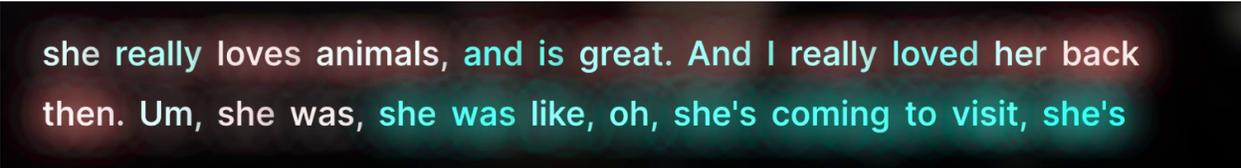
The BWS portion of the test was similar to that of Phase 1. Excerpts were slightly longer ($\mu = 27$ s, with $\sigma = 5$ s), and accounting for how the comparisons themselves were more complex – two styles per video, with subtle differences between some combinations – we reduced the number of videos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

After completing the BWS/videos portion of Phase 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 [201], this data, along with notes taken during this second phase of the study, and the open-ended data collected in Phase 1, was separately analyzed and clustered by two authors to answer RQ4 regarding the factors that influence DHH viewers' preference for specific caption text styles.

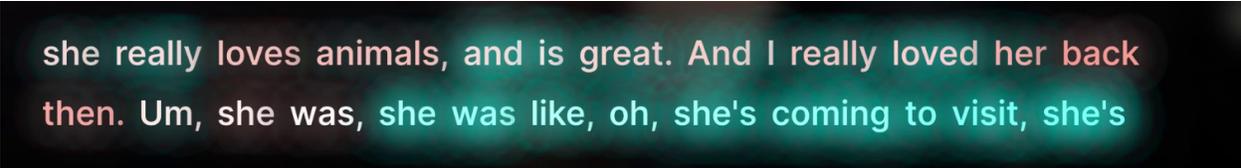
6.3.1 *Findings from Phase 2*

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 Deaf or Hard-of-Hearing. For Phase 2 we recruited a total of 11 participants, 8 of which identified as female and 3 as male, 5 of who identified as Deaf and 6 as Hard-of-Hearing, with a mean age of 28.5 years ($\sigma = 9.7$).



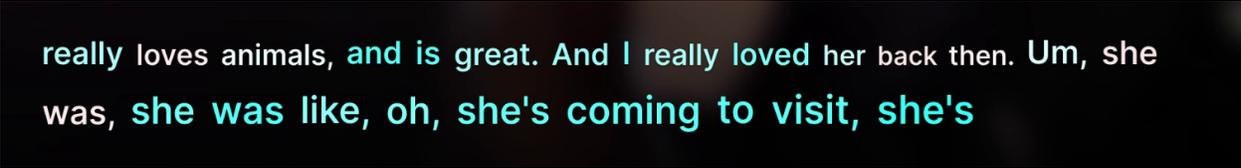
she really loves animals, and is great. And I really loved her back then. Um, she was, she was like, oh, she's coming to visit, she's

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



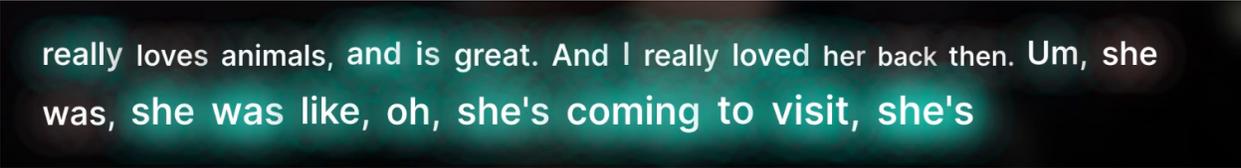
she really loves animals, and is great. And I really loved her back then. Um, she was, she was like, oh, she's coming to visit, she's

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



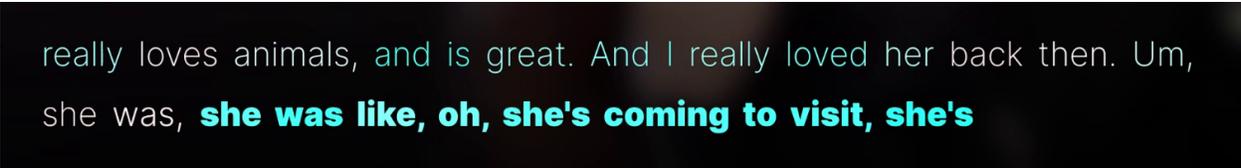
really loves animals, and is great. And I really loved her back then. Um, she was, she was like, oh, she's coming to visit, she's

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



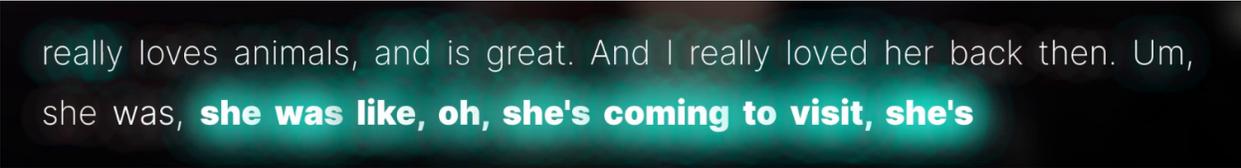
really loves animals, and is great. And I really loved her back then. Um, she was, she was like, oh, she's coming to visit, she's

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



really loves animals, and is great. And I really loved her back then. Um, she was, she was like, oh, she's coming to visit, she's

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



really loves animals, and is great. And I really loved her back then. Um, she was, she was like, oh, she's coming to visit, she's

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

Figure 6.7: The six caption styles used in Phase 2 of the evaluation. All images are screenshots from one of the videos used as stimuli.

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

Table 6.2: 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.

6.3.1.1 *Caption Style Reference Rankings*

Phase 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 6.2 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 $p < 0.051$.

Running the pairings through the TrueSkill, again with the same parameters as used in Phase 1, gave us the following values: **Font-color with font-weight** ($\mu = 26.5$, $\sigma = 0.8$), **font-color with font-size** ($\mu = 26.2$, $\sigma = 0.8$), **font-color with shadow-color** ($\mu = 25.8$, $\sigma = 0.8$), **shadow-color with font-color** ($\mu = 25.2$, $\sigma = 0.8$), **shadow-color with font-weight** ($\mu = 23.4$, $\sigma = 0.8$), and **shadow-color with font-size** ($\mu = 23.1$, $\sigma = 0.8$).

6.3.1.2 *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 –*

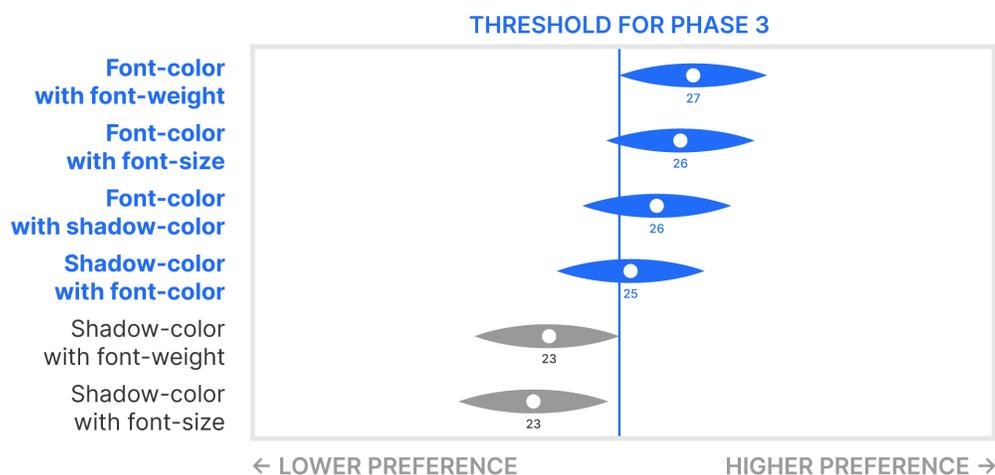


Figure 6.8: Relative strength and confidence interval for each caption style tested in Phase 2. Style names have the first style depicting valence, and the second arousal. Styles in bold blue font have their TrueSkill μ value overlapping the top choice when considering the 95% confidence interval. As with Phase 1 of this study, the initial μ value was set at 25, meaning that styles that finished the matches above that value gained skill points.

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.'*

6.4 Phase 3: Subjective and Objective Performance of the Valence-Arousal Combined Caption Styles Against a Neutral Baseline

At this point, Phases 1 and 2 had given us two important insights. First, they allowed us to narrow an initial set of 72 possible caption style combinations⁷ 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 Phase 3, focusing on two complementary aspects: participants' ability to recognize emotions from captioned videos, and their subjective evaluation of how the different styles measured up against the four criteria established in Phases 1–2.

6.4.1 Methods

6.4.1.1 Stimuli and Experimental Design

We once again used 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

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

in the following conditions: (1) baseline, featuring non-stylized captions; (2) font-color with font-weight (Figure 6.7e); (3) font-color with font-size (Figure 6.7c); (4) font-color with shadow-color (Figure 6.7a); and (5) shadow-color with font-color (Figure 6.7b).

One of the factors outlined in answering RQ4 was that affective captions should have a CLEAR EMOTIONAL REPRESENTATION, done so INTUITIVELY. *Intuitive* is a fraught term in HCI [199], 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 phases of this study, 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.

6.4.1.2 *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 [53, 99]. 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 [114], by comparing the novel caption style conditions with a 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-Šidák corrections.

The second approach, previously explored by Hassan et al. [85], involved having participants annotate single words from a captioned video. We expanded on this method by having participants annotate four distinct ten-word⁸ 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.

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 [181]. 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.*, [182, 183]. 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 its limit of only considering two variables at a time [195], *i.e.*, correlating ground-

⁸The 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.

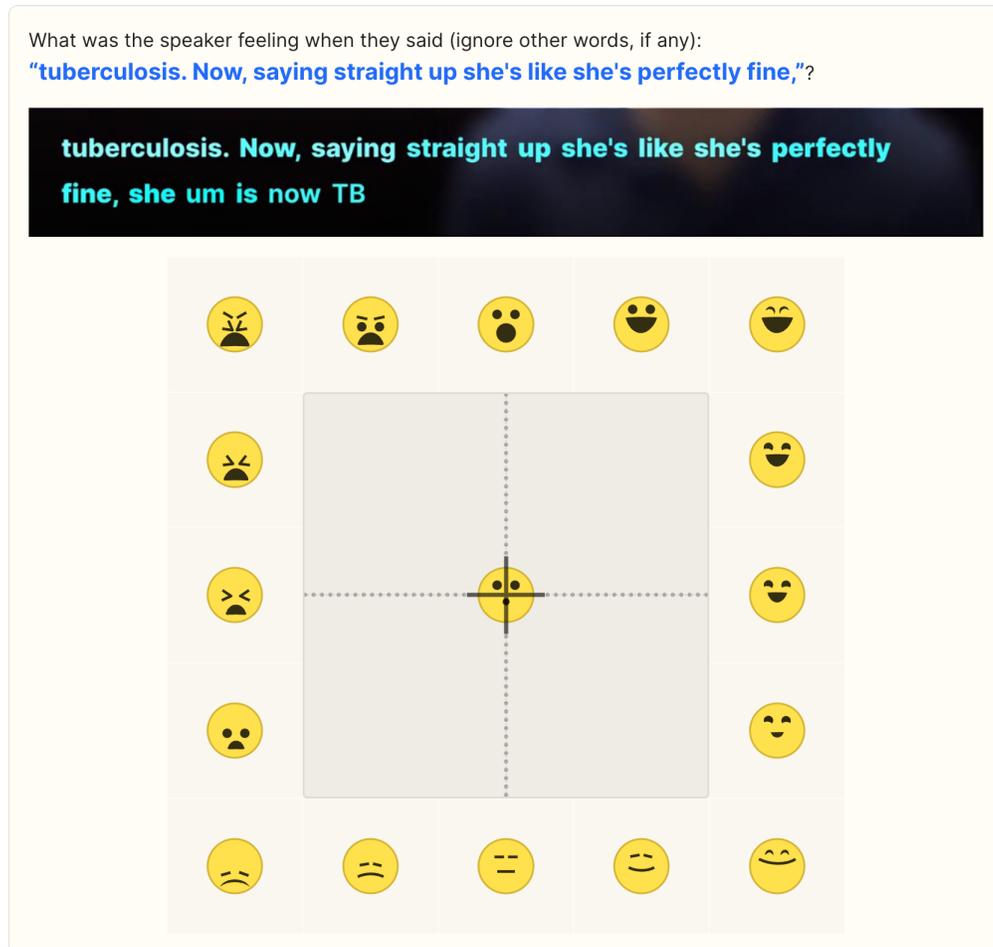


Figure 6.9: Our EmojiGrid implementation, with figures from Toet [180]. In this example, participants had to label an image that – as we know beforehand, but they had to figure out – has valence, depicted by font-color, and arousal, depicted by font-weight, both placed on the top-right quadrant.

truth valence versus participants' valence while ignoring the corresponding pair of arousal values and vice-versa.

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 [178]. 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.

6.4.1.3 *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. [99], 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 [81, 96]. 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*.

6.4.2 *Findings from Phase 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 Deaf or Hard-of-Hearing. For Phase 3 we recruited a total of 18 participants, 10 of which identified as female and 8 as male, 11 of who identified as 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.

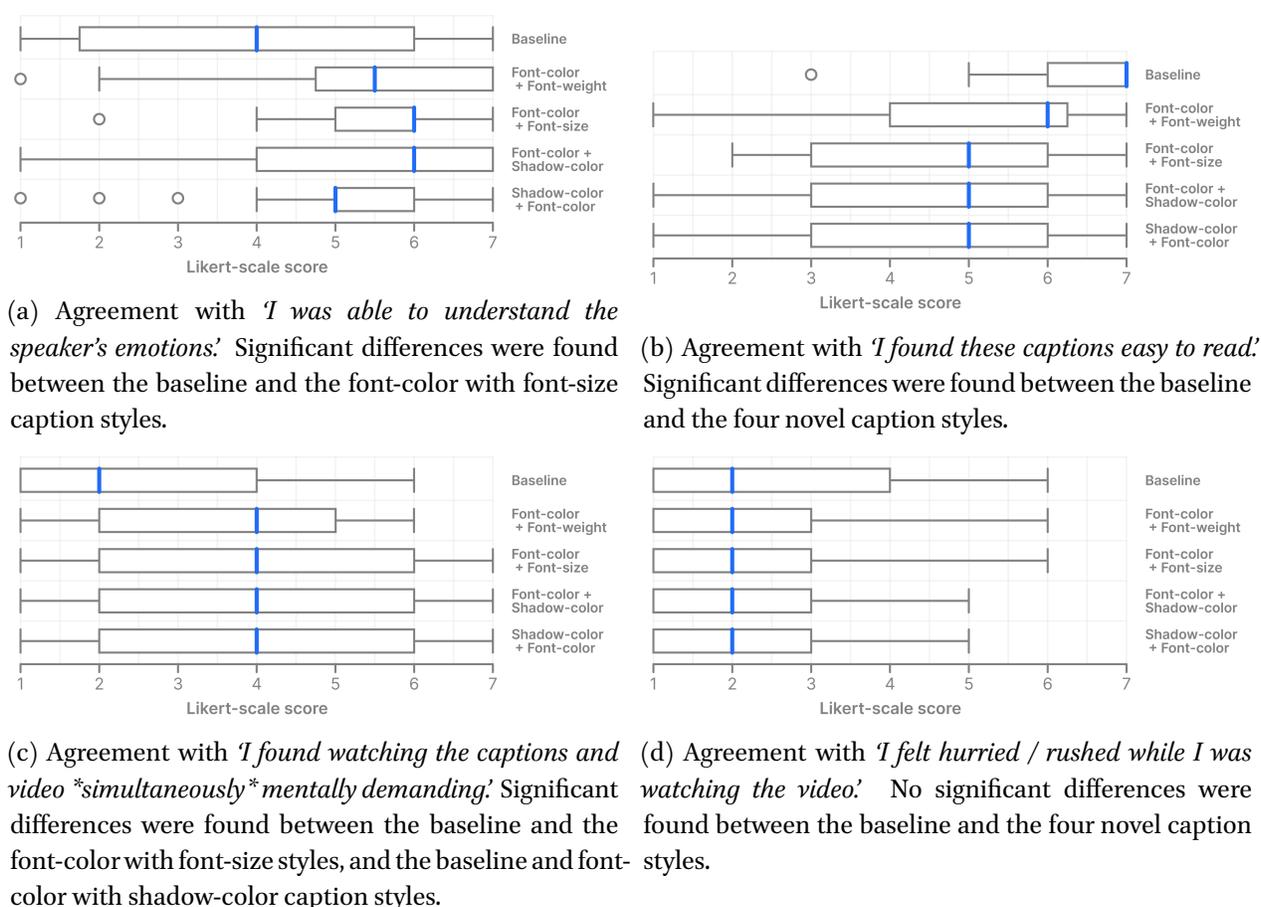


Figure 6.10: Box-whisker plots with spread of answers between the five conditions for different Likert scales.

6.4.2.1 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 6.10a, were: 4.0 for the baseline condition; 5.5 for font-color with font-weight; 6.0 for font-color with font-size, with significant difference versus the baseline ($U = 385.0$, $p < 0.05$, *medium effect*)⁹; 6.0 for font-color with shadow-color; and 5.0 for shadow-color with font-color.

We calculated the distance correlation between ground-truth and participant-provided coordinates using Ramos-Carreño and Torrecilla's implementation [153, 154]. The resulting correlation coefficients are presented in Table 6.3. Notably, font-color with font-weight and font-color with font-size exhibited a sig-

⁹P-values presented adjusted using Holm-Šidák corrections.

nificant 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
<i>Baseline</i>	0.14	0.09	0.07
<i>font-color</i> + <i>font-weight</i>	0.14	0.32***	0.21***
<i>font-color</i> + <i>font-size</i>	0.10	0.23**	0.14*
<i>font-color</i> + <i>shadow-color</i>	0.10	0.10	0.06
<i>shadow-color</i> + <i>font-color</i>	0.10	0.10	0.07

Table 6.3: 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.

6.4.2.2 Ease of Reading and Processing

Median responses for agreement with the *I found these captions easy to read* statement, shown in Figure 6.10b, 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 6.10c, 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 6.10d.

6.5 Discussion

While prior research had investigated the benefits of affective captions and their effectiveness at conveying emotions [49, 51, 53, 85, 99], 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 (RQ3.A and RQ3.B), the factors that influence these choices (RQ3.C), and how well the preferred styles performed under a set of meaningful objective (RQ3.D) and subjective (RQ3.E) quality criteria.

6.5.1 Caption-Style Preferences (RQ3.A & RQ3.B)

Findings from Phase 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 RQ3.A, the outcome of Phase 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 Phase 1's styles for Phase 2 yielded six new styles. Answering RQ3.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 [53, 85], while also showing that many typographic parameters considered for prosodic captions were not as effective [38, 51, 52, 205].

6.5.2 *Factors Influencing Participants' Choices of Caption Styles (RQ3.C)*

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. [85], 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 RQ3.C, 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.

6.5.3 *Objective Measures of Performance (RQ3.D)*

Table 6.3 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.D, these results show that, although participants did not significantly prefer one of the four caption styles over the others (Phases 1–2), the objective metrics from the emotion-recognition task in Phase 3 told a different story. Specifically, correlation coefficients between participants' responses and the ground-truth valence/arousal ratings (Table 6.3) were significantly higher for the two styles combining font-color (valence) with either font-weight or font-size (arousal) than for the two styles that combined font-color with shadow-color. This indicates that the former two styles enabled

participants to more accurately identify the intended emotions, despite not being more preferred in earlier phases.

6.5.4 *Subjective Measures of Performance (RQ3.E)*

We also measured how much participants *felt* each caption style helped their understanding of the speaker's emotions (see Figure 6.10a). 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 6.10c). 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 Phase 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 6.10b).

6.5.5 *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 is not 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.

6.6 *Limitations & Future Work*

The emotional richness of the SEND videos used in this study 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 [150], accompanied by modules for word-level timestamping [117], voice activity detection [204], and emotion recognition [190]. 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 [95]. 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 [79, 99], but differences exist in perception, such as a smaller legibility drop for Hangul [53, 99]. 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 [37, 191], 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.*, [196]. 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.

6.7 *Conclusion*

In Phase 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 Phase 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 was able to provide a CLEAR REPRESENTATION OF EMOTIONS.

In Phase 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 Phase 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.

PART III

Using Haptic Feedback to Convey a Speaker's Arousal Levels

Includes Study 4, 5, and 6

CHAPTER 7

Introduction

As we saw in Study 2, for applications aimed at making speech-modulated captions that are useful to DHH individuals, depicting some features from speech *but not others* can lead to better outcomes: DHH folks preferred text that could convey a speaker's emotions over text that depicted the actual sound of their voices, *i.e.*, *affective* captions were preferred captions over those that displayed *prosodic* features such as loudness, pitch, and rhythm.

This finding has led us and other researchers to explore the design of affective captions [55, 85, 99]. In Study 3, we saw a clear preference for color-coded captions indicating a speaker's valence (positive or negative tone). However, preferences for depicting a speaker's arousal level were less clear. While the evidence suggests font-weight or size would be viable typographic parameters, these styles had only a slight edge in user preference compared to the evaluated alternatives.

Looking at Study 3, we wondered: could there be inherent differences in how specific aspects of emotions are perceived? If visual cues effectively convey valence, perhaps arousal requires a different, non-visual communication channel. In particular, could arousal be conveyed through haptic feedback? Our approach aligns with findings by Akshita et al. [4], who showed that in multimodal visual-haptic stimuli, the visual channel predominantly shaped perceptions of valence (analogous to our use of caption modulations), while the haptic channel influenced perceptions of arousal (not unlike our encoding of a speaker's arousal through vibrations). Building on this principle, Studies 4–6 explored whether haptic feedback could serve as that complementary channel: identifying how best to encode arousal through haptics (Study 4), assessing the effects of combining vibrations with visual modulations on engagement (Study 5), and testing how well participants could decode our proposed haptic-arousal mapping (Study 6).

7.1 *Study 4*

In the first part, we searched for the combination of haptic patterns that participants felt worked best to represent arousal. Prior work suggests varying the intensity of the signal, but gives little guidance as to what its rhythm and frequency should be so that both the message is clear and the vibrations are comfortable in feeling. Thus, our first research question looked at these two parameters, hoping to bring light to the design space of haptic feedback signals coupled with affective captions:

RQ4 What combination of a rhythmic pattern and frequency, presented as haptic feedback, is perceived as the most effective and comfortable for conveying a speaker's arousal levels, as judged by DHH individuals?

Here, and across all conditions, the haptic signal's amplitude was modulated to convey a speaker's arousal levels – *e.g.*, high arousal presented as strong vibrations, low arousal as weak vibrations, and so on – with varying rhythmic patterns and frequencies left for participants to judge.

7.2 *Study 5*

Following this initial research question, we turned to investigating whether this signal actually complements a visual depiction of arousal, as previously explored in Study 3, or, alternatively, if it can serve as a standalone representation, or, ultimately, whether it does not actively help:

RQ5 How do haptic feedback and typographic modulations, used alone or in combination, influence arousal depiction and narrative engagement for DHH individuals when compared to a baseline comprised of standard, neutral captions?

To answer RQ5, we measured participants' *narrative engagement* with audio-visual content. In other words, do these different conditions help make content more engaging for DHH viewers? As a means of comparison, we also included a neutral caption style as a baseline that had no affective cues, visual or haptic.

7.3 Study 6

Answering RQ5, we saw that, indeed, combining haptics and visuals improved DHH participants' *engagement* with audio-visual media – a satisfying response to (at least part of) the *dull-ambiguity* our interviewees in Chapter 2 reported with current captioning systems. This leaves open whether adding haptics to convey arousal affects how *ambiguous* this dimension is. While in RQ3.D the visual modality of affective captioning outperformed a conventional baseline in helping participants discern emotions in speech, haptics could disrupt this – a possible tradeoff that, if confirmed, might lessen its engagement benefits.

As such, in our follow-up Study 6, we echoed RQ3.D from Chapter 6 in examining whether this newly established haptic pattern improves recognition of a speaker's arousal levels. Because interaction effects between visuals and haptics were possible, we tested conditions that both isolated and combined these modalities, comparing them against a neutral baseline. We sought to answer:

RQ6 How well do modulations of haptics and typography map to how DHH participants perceive a speaker's arousal levels?

CHAPTER 8

Background and Related Work

In this section, we will examine papers that explore the use of haptics to enhance the perception of speech and music for DHH individuals, as well as their potential for conveying emotions. Although there is currently no specific research investigating the combined effects of typographic modulation and haptics in depicting affective states, the findings from these individual haptic-based studies inform our approach and shed light on the subject.

8.0.1 *Interpretation of Haptic Signals*

In their survey of tangible user interfaces, Zhou et al. [207] identify a tension between haptic devices designed for clarity and those that invite ambiguity. Rather than a flaw, the latter can be seen as an invitation to sense-making [72]. Although much affective computing treats emotion as quantifiable [147], we highlight an alternative view: meaning is enacted in context [29]; more than just form, emotional interpretation relies on relational, situational, and cultural embeddings. The same cue can thus read differently depending on context – *e.g.*, a strong vibration echoing speech might suggest anger, or simply vocal effort in a noisy nightclub – and, as seen by Sundström et al. [176], open-ended designs can lead users to invent their own encodings of affect.

In haptics, this open-endedness reflects embodied associations: because touch mediates our relation to the world, prior tactile experiences can bias how we perceive tangible interfaces. Studies show that haptics can shape perception across modalities, even when the signals are unrelated. Gatti et al. [71], for instance, found that haptic feedback in a pointing device altered how participants rated arousal levels in a set of images, while Salminen et al. [163] showed that haptic echoes of a speech signal's amplitude heightened how arousal was perceived. Likewise, Ackerman et al. [1] demonstrated that incidental sen-

sations – such as holding a heavy clipboard or sitting in a hard surface – unconsciously biased social judgments, with weight evoking seriousness (e.g., judging a job candidate) and hardness evoking rigidity and stability (e.g., during negotiations). Such effects support the view that sensorimotor experience scaffolds conceptual knowledge, with later tactile encounters drawing on these embodied metaphors [1]. This perspective emphasizes that haptic signals are not merely delivered but *interpreted* in context. With this in mind, we next consider work that approaches haptics as a channel for translating sound and emotion in more structured ways.

8.1 *Haptics as Sound/Emotion Translating Channel*

Haptic technologies apply physical stimuli, such as forces, heat, or vibrations, to a user's body, stimulating tactile sensations [119]. These technologies have a wide range of areas, from tactile feedback systems that help pilots maintain safe flight parameters [165] to improving medical training [5]. Within the HCI and accessibility communities, there has been a growing interest in the use of haptics to convey aspects of sound, such as speech [67], environmental sounds [91], and music [66, 125], for DHH individuals, as they can transmit information without overloading the visual channel.

In a study surveying DHH people's preferences for sound awareness technologies, smartwatches came on top [64]. Because of their mainstream appeal, they can avoid the stigma often associated with dedicated assistive devices [171]. Additionally, the haptic feedback provided by smartwatches can effectively complement visual information displayed elsewhere, a concept we explored in our study that was also demonstrated by Goodman et al. [75].

Other haptic devices have also been used to complement visual information. Weisenberger et al. [197], for example, found that translating sound into tactile signals speech reading accuracy for DHH people. This was echoed by Fletcher et al. [67], who showed that haptic feedback can improve speech intelligibility performance for cochlear implant users, particularly after a period of training.

In another study, Wang et al. [196] explored translating speech sounds into haptic-feedback, helping their DHH participants differentiate between speakers and intuit their moods. Interestingly, they achieved this using a simple setup: a voice coil actuator placed in a 3D-printed wrist-worn casing, driven by a 3W power amp – cheap and readily available components. For our studies, we employed a similar setup (see section 9.2 for more details).

In essence, these methods exemplify the concept of *sensory substitution*, where one sense is supplemented with information that would be typically gathered by another [120]. In this context, sound elements are often translated into haptic feedback. While haptics can include various touch sensations like pressure, temperature, shape, and texture, the examples mentioned primarily use *vibrations* which, according to Flores Ramones and del Rio-Guerra, share qualities with sound, such as frequency, amplitude, and duration [68]. However, directly mapping sound to haptics in a 1 : 1 manner, though feasible, presents several challenges.

For one, there are significant differences in frequency response curves for sound, *e.g.*, the Fletcher-Munson curves for hearing [65], versus that of touch at different parts of the skin, *e.g.*, [44, 187]. Privacy and comfort concerns also arise, particularly when music or human speech is used as direct input to vibrotactile haptics systems. Such signals contain frequencies in the audible range (approximately 40 Hz – 18 kHz), creating audible sounds through the haptic system that can compromise privacy. These can also cause discomfort due to tingling sensations from vibrotactile stimulation at frequencies above 200 Hz [131]. Verrillo [187], for instance, found that the *sensational quality* of vibrations below 100 Hz differ from those at higher frequencies, with the former producing a *buzz*-like sensation and the latter a smoother one.

Other researchers have investigated how haptics can convey information that may have no direct real-world correlates. Ternes and MacLean [179], for instance, examined varying patterns of amplitude, frequency, and rhythm to create 84 unique haptic *icons* that developers and designers can use to convey information. Amplitude was identified as the most strongly perceived differentiating factor. This haptic ‘vocabulary’ was further explored by Seifi and MacLean [167], who found that participants assigned different affective categories to different stimuli, *e.g.*, long vibrations were perceived as pleasant, while repeated short vibrations were felt to be alarming and unpleasant. These explorations of the design space of haptic feedback inform the first phase of our study, detailed in subsection 10.1.

Akshita et al. [4] showed that parameters of a synthetic haptic signal with no external correlate can intensify an individual’s emotional response to images, particularly arousal – supporting our approach of combining affective captions with haptic feedback.

CHAPTER 9

System Design

This section describes the design of the haptic-captioning system used in Studies 4–6. Although there were differences in the setup for each individual study, they all employed the same core functionality, which is presented in full here. The system described here shares features with the prototypes presented in previous chapters, but it was developed from scratch to better integrate with ASR transcription and generation of haptic stimuli, which were needed for the current experiments.

First, we present our pipeline to process each video’s audio files, obtaining speech transcriptions with corresponding affective features (section 9.1). Second, we go over how we defined the haptic signal that echoed speech arousal, and how it was used to drive a wrist-worn haptic device (section 9.2). Third, we discuss how we implemented the visuals applied to the typography of the captions used in the two studies (section 9.3).

9.1 *Transcription and Emotion Recognition of a Speech Signal*

All videos were transcribed using OpenAI’s Whisper speech recognition model [150] with word-level timestamping [117]. The voice activity detection (VAD) flag was enabled to improve transcription when background noises were present.

As before, we employed the circumplex *dimensional* model of emotions [160] using Wagner et al.’s open-source toolkit [190] for emotion recognition, configured to output valence and arousal levels for each individual word. The predicted values were included as metadata added to each word of a WebVTT caption file [49].

9.2 Using a Haptic Signal to Convey a Speaker's Arousal Levels

The same arousal information that can be used to modulate the visual attribute in the typography of captions can also be used to modulate a haptic signal. In fact, part of our contribution is the novel approach we present to do so, *i.e.*, the way that the intensity of a haptic signal can be modulated using these values so that a viewer has a sense of how excited or calm the emotions in a speaker's voice are. Here, a strong vibration would follow an excited emotion, while a calm emotion would be echoed by a fainter vibration.

The choice of *intensity* as the *modulated dimension* comes from Ternes and MacLean [179], who in their study of haptic icons found that amplitude was the most distinctly perceived differentiating factor. In other words, changes to it were easier to perceive than changes to the two other dimensions that, as per Akshita et al. [4], comprise a haptic signal: frequency and rhythm.¹

To drive this signal, a physical device was needed. Given Findlater et al.'s finding that smartwatches were the preferred form factor for sound awareness tech [64], we followed Wang et al.'s haptic captioning study [196] and used Acouve's Vp2 Vibro-Transducer,² a simple voice coil driven by Techtile Toolkit's power amplifier [133]. It converts audio signals sourced from a laptop's audio jack into haptic vibrations. The device was housed inside a 3D-printed casing, which could be attached to participants' wrists using a velcro band, as seen in Figure 9.1.

To generate the audio files driving the haptic patterns, we wrote a ChuckK language script [194] that converted arousal values encoded in the caption file into a sound signal to be played alongside the video. ChuckK operates on a *strongly-timed* paradigm, which guarantees precise temporal accuracy in the programmed sounds down to the sample level. This ensured that the generated sounds remained synchronized with the video.³ To allow for the playback of both the original video audio and the haptic-generating sound files, we used a stereo sound signal where each one of the two channels corresponded to a distinct output. A stereo splitter cable was used to route the outputs to their respective devices.

¹Akshita et al. also lists waveform as a component of haptic signals, but their tests saw evidence that it does not influence the perception of arousal, prompting us to simplify our approach by utilizing sine waves exclusively [4].

²<https://www.acouve-lab.com/products>

³The actual implementation was done using the WebChuck toolkit [136], which runs in a web environment and, as such, could be integrated into the same web-based script described in de Lacerda Pataca [49] that we used to generate the visually-modulated affective captions.



Figure 9.1: To watch videos, participants would strap the voice coil to their arm, with the device face-down against the inside of their wrist. A laptop would drive both the haptic signals and an external speaker, that played the original sounds coming from the videos.

I should have ordered a decaf or a tea. No, it's fine. I've made a decision. I can have as much caffeine as I want. And sugar.

Figure 9.2: Example of how typographic attributes can be modulated to convey a speaker's valence and arousal levels. Here, valence is represented by font-color, with red indicating that the first sentence was said in a negative tone, transitioning to a more neutral and lightly positive tone as they say 'much caffeine.' Arousal is shown by changes to font-weight (thickness), reaching its highest when they say 'it's fine.'

9.3 *Typographic Representations of Valence and Arousal Levels*

Typographic modulations were used in both phases of the study to convey speech features and, in some conditions, were combined with haptic feedback. Our approach was similar to that described in 6.2.1.1.2; see that section for further details. Of note, the speech → typography mappings we used were based on the top-performing choices from Chapter 6, namely, font-color mapped to valence, and font-weight to arousal. An example of this font-color / font-weight modulation is shown in Figure 9.2.

CHAPTER 10

Study 4: Using Haptic Feedback to Convey a Speaker's Arousal Levels¹

Building on prior work that has *intensity* as the primary haptic dimension for communicating differences in arousal, Study 4 aimed at experimentally determining which *frequency* and *rhythmic* properties of this signal are perceived as effective and comfortable for doing so.

10.1 *Defining the Different Haptic Patterns*

10.1.1 *Rhythm*

Rhythm refers to the variations in the haptic signal over time. Our goal is to determine whether the vibration should be continuous throughout a speaker's utterance, include pauses to emphasize individual words, or have its own independent rhythm. To investigate this, we selected three distinctly different

¹This study was part of a joint project between myself, Dr. Saad Hassan, assistant professor at Tulane University, Lloyd May, Ph.D. student at Stanford University, Michelle M. Olson and Toni D'Aurio, graduate students at RIT, and my co-advisors, Dr. Roshan L. Peiris and Dr. Matt Huenerfauth. I led the study design, stimuli creation, data analysis, and writing of the paper, which was published at the ACM CHI'25 conference [54].

rhythmic patterns from the set defined by Ternes and MacLean [179]. These patterns are listed below and schematically represented in Figure 10.1:

LP A LONG PULSE, vibrating for the whole duration² of the word (Figure 10.1a);

SSP A SINGLE SHORT PULSE lasting for two-thirds of the duration of the word, with a one-third silence at its end (Figure 10.1b);

MSP A series of MULTIPLE SHORT PULSES with a fixed duration.³ The number of pulses will be proportional to the duration of the word itself (Figure 10.1c).

10.1.2 Frequency

Along with *rhythm*, Akshita et al. [4] conceptualizes *frequency* as a defining property of a haptic signal. In essence, this is the rate at which the haptic signal oscillates, typically measured in Hertz (Hz). Frequency determines the pitch of the vibration, a property that is related to perceptual qualities of the haptic feedback signal.

Following how Ævarsson et al. define ranges of maximum sensitivity at the wrist [2], we defined two frequency levels: a low tier, at 75 Hz, and a high tier, at 250 Hz.

From previous perception literature [2], we learn that on the glabrous (non-hairy) skin of the wrist, the threshold of detection of a vibrotactile signal of 250 Hz is approximately 10 dB higher than for a 75 Hz one. Although the perceptual metric gauged detection threshold and not equal intensity, for the purposes of coarse calibration, we believe this 10 dB perceptual difference offset is sufficient as we could not locate research that established perceptually equivalent intensity levels across frequencies of vibrotactile stimulation on the wrist.

We performed frequency calibration of the hardware through recordings using a piezoelectric surface microphone with the hardware freely vibrating and under a 2 kg load, approximating the loading condition of being strapped comfortably to a participant's wrist. In both situations, the amplitude of the measured

²To avoid pops on the haptic device's speaker, we apply fade-in and fade-out for the envelopes for attack and release of the signal, each lasting either 1/40th of the duration of the word or, if the word is too short, 12.5 ms or 1/2 of the word, whichever is shorter.

³The duration of these pulses follows Seifi and MacLean [167], whose fast-pulse rhythmic pattern uses 62.5 ms pulses (1/16 s).

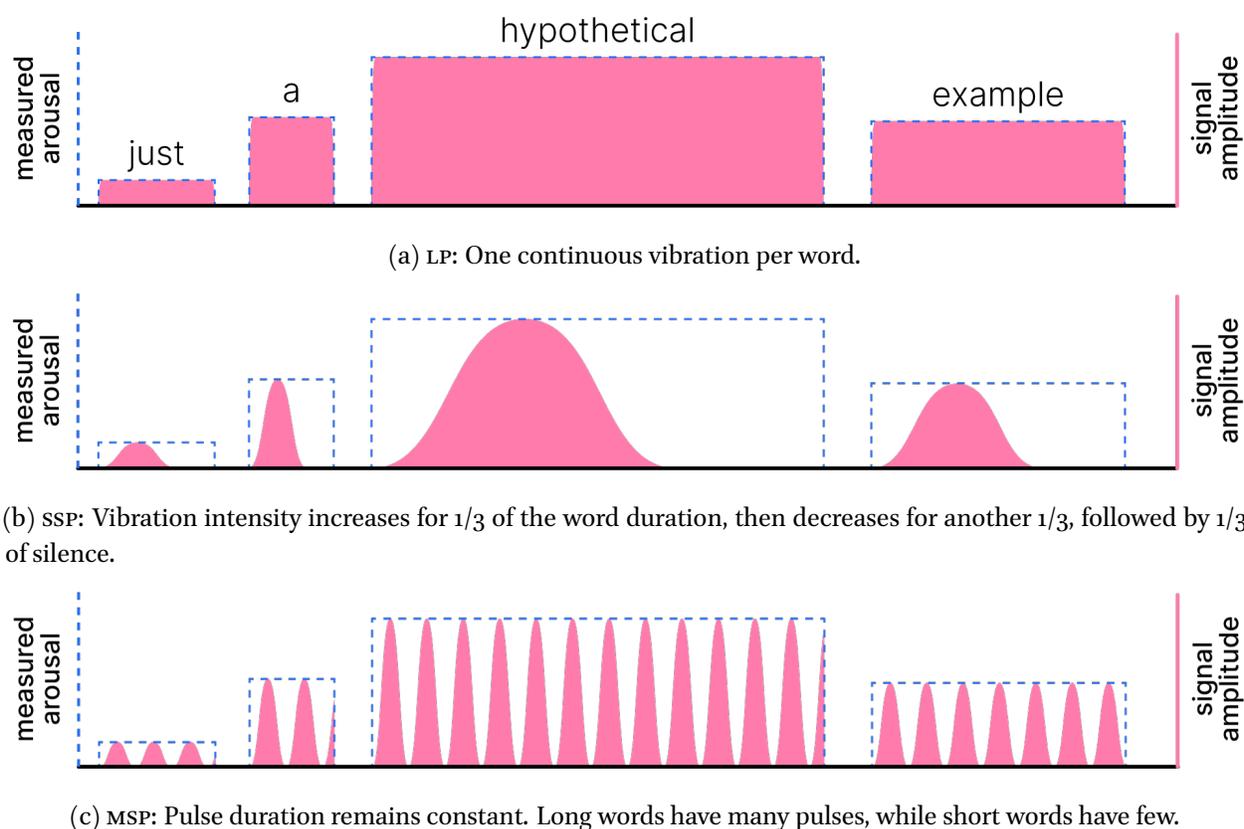


Figure 10.1: These charts illustrate how three haptic signal configurations (y-axis, right) respond to changing arousal levels (y-axis, left) over time (x-axis; aligned to word onsets). The dashed-blue lines indicate the predicted arousal values for each word, while the shaded-pink areas show the duration and intensity of each corresponding haptic vibration. The phrase 'just a hypothetical example' is spoken with increasing arousal from 'just' to 'hypothetical,' then decreasing for 'example.'

FREQUENCY	RHYTHMIC PATTERN		
	<i>Long Pulse</i>	<i>Single Short Pulse</i>	<i>Multiple Short Pulses</i>
75 Hz	LP _{75 Hz}	SSP _{75 Hz}	MSP _{75 Hz}
250 Hz	LP _{250 Hz}	SSP _{250 Hz}	MSP _{250 Hz}

Table 10.1: The six haptic conditions evaluated in Study 4.

75 Hz waveform was approximately 4 dB lower than the 250 Hz waveform. Because the perceptual deficit at 250 Hz is 10 dB yet the actuator already delivers a 4 dB hardware boost at that frequency, we added only +6 dB to the 250 Hz drive level ($10 - 4 = 6$) to approximate equal perceived intensity while keeping overall intensity constant as a potential confound.

The two frequencies, combined with the three rhythmic patterns, gave us the six total conditions, or haptic patterns, evaluated in this first study and presented in Table 10.1.

10.2 *Experimental Procedure*

Participants were recruited by sending out IRB-approved ads to social network groups and university-related student groups. Participants qualified to participate in this experiment if they identified as Deaf or Hard-of-Hearing. For Study 4 we recruited a total of 16 participants, 9 of which identified as female and 7 as male, 11 of which identified as Deaf and 5 as Hard-of-Hearing, with a mean age of 27.1 years ($\sigma = 8.9$). A compensation of \$40 was offered.

The study was conducted in person. Upon arrival, participants met with an ASL-native research assistant who explained the study. After agreeing to take part in the study, participants were assisted in attaching a haptic device to their non-dominant hand. A test haptic signal was played to ensure the device's intensity was comfortable. Once this setup was complete, participants began the study, which was conducted through an interactive website.

The website was developed using jsPsych [56], and is shown in Figure 10.2. The number of stimuli shown for each participant echoed the study described in Chapter 6, *i.e.*, 10 rounds, each with a different video, with four conditions tested per round, and again employing a best-worst scaling setup. We counterbal-

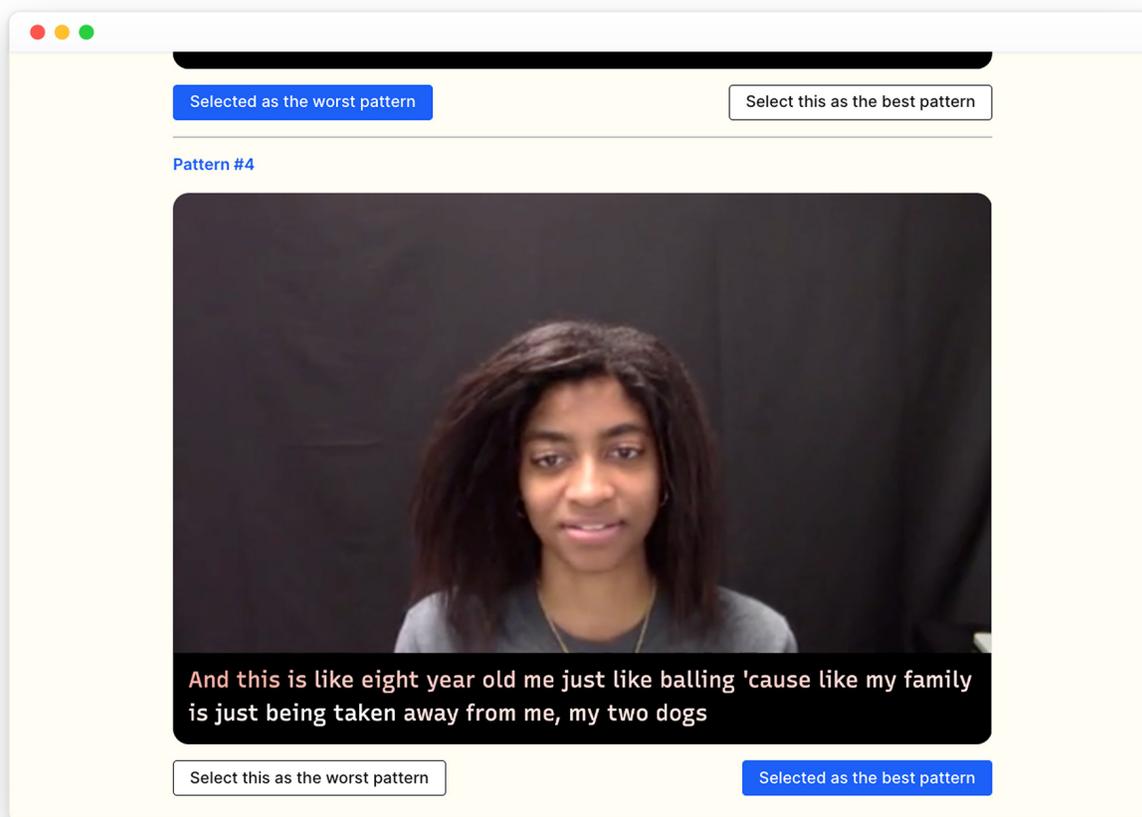


Figure 10.2: Screenshot of the test platform.

anced video and condition order. As in studies 2 and 3, for Study 4 videos were sourced from the Stanford Emotional Narratives Dataset [143].

Participants were asked to watch each short video excerpt in its entirety at least once, with the option of rewatching them as needed. Using keyboard or mouse, participants were asked to ‘Select the vibration patterns that you believe best and worst reflect the intensity of [the speaker’s] emotions.’

Finally, participants answered the following questions: ‘Did the vibrations influence how you understood what the speaker was saying in the different versions of the same video? If yes, could you provide further details?’, ‘What aspects of the best vibration patterns do you think worked well?’, and ‘What issues did you encounter with the worst vibration patterns?’

10.2.1 *Analysis plan*

As mentioned, we adopted a best-worst scaling (BWS) methodology, which allowed us to measure participants' preferences towards the six haptic patterns by establishing a simple criterion – which patterns best and worst convey the intensity of the speaker's emotions? – and prompting participants to judge which stimuli are the *best* and *worst* examples of it. Following Study 3 and recent examples in HCI research [146], we adopted Herbrich et al.'s TrueSkill implementation of an ELO-rating system [86], with Clark et al.'s recommendation to average rankings across randomly ordered iterations to address the order-independence of BWS tuples.

10.3 *Findings from Study 4*

10.3.1 *Haptic pattern rankings*

Study 4 had 16 participants evaluating 4 videos per round for 10 rounds. A 4-way BWS generates 5 data pairs, so $16 \times 10 \times 5 = 800$ pairwise comparisons. Table 10.2 shows the results from the study, including both the raw answers – *i.e.*, what participants explicitly chose (or 'N/A', for the times a pattern was shown but was not explicitly chosen as either the best or worst option) – and the choices implied by leveraging the BWS setup.

The TrueSkill values – where higher values correspond to higher levels of preference – for the LP rhythmic pattern (the long pulse) with 75 Hz and 250 Hz were, respectively, $\mu = 23.8$, $\sigma = 0.8$, and $\mu = 22.1$, $\sigma = 0.8$. Values for the SSP rhythmic pattern (the shorter, single pulse) with 75 Hz and 250 Hz were, respectively, $\mu = 29.6$, $\sigma = 0.8$, and $\mu = 27.7$, $\sigma = 0.8$. Lastly, values for the MSP rhythmic pattern (the multiple fixed-duration pulses) with 75 Hz and 250 Hz were, respectively, $\mu = 24.3$, $\sigma = 0.8$, and $\mu = 22.4$, $\sigma = 0.8$. These values are also shown in Figure 10.3. Note that before processing participants' preferences, each haptic pattern was initialized with a skill of 25.⁴

⁴TrueSkill parameters set at their default values of $\mu = 25$, $\sigma = \mu/3$, $\beta = \sigma/2$, and $\tau = \sigma/100$.

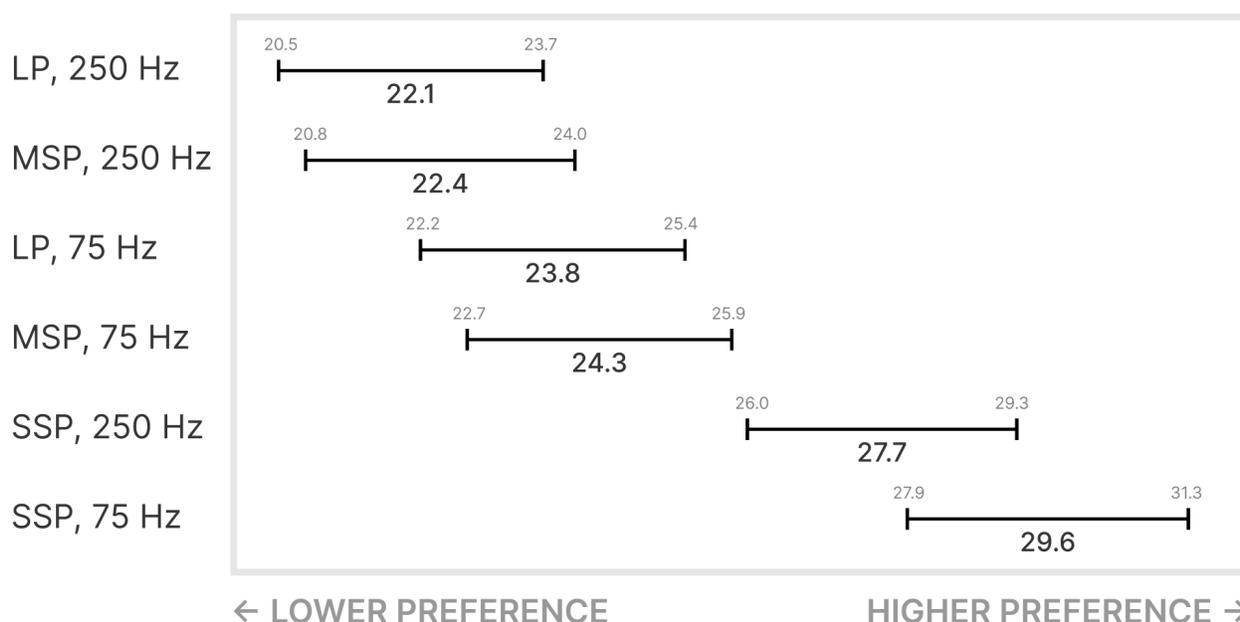


Figure 10.3: Final TrueSkill rankings of the six haptic patterns. LP represents the long pulse; SSP, the single short pulse; MSP, the multiple short pulses. These are combined with two frequencies, 75 Hz and 250 Hz. The skill is shown below each line, with its 95% confidence range shown above.

<i>Haptic pattern</i>	RAW ANSWERS			IMPLIED ANSWERS	
	WON	LOST	N/A	WINS	LOSSES
<i>LP, 250 Hz</i>	14%	48%	38%	33%	67%
<i>MSP, 250 Hz</i>	10%	36%	54%	37%	63%
<i>LP, 75 Hz</i>	17%	31%	52%	43%	57%
<i>MSP, 75 Hz</i>	17%	24%	59%	47%	53%
<i>SSP, 250 Hz</i>	38%	9%	53%	64%	36%
<i>SSP, 75 Hz</i>	54%	3%	43%	76%	24%

Table 10.2: Raw and implied (as per the BWS method) results for each one of the six haptic patterns. In the raw results columns, choosing a pattern 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 pattern was shown in a round but was not marked as best or worst option. The ordering of the table follows the patterns' ascending top-to-bottom TrueSkill values, also shown in Figure 10.3. LP represents the long pulse; SSP, the single short pulse; MSP, the multiple short pulses.

10.3.2 *Open-Ended Comments*

Participants shared their thoughts on what worked well and what did not with the different haptic patterns, as well as more general feedback on using vibrations to represent speakers' arousal levels. Below, we present a summary of these ideas, supported by participant quotes. Where needed, quotes were edited for clarity.

Some participants felt that different emotions had a different tactile feel. For P1, happy emotions were 'more peppy or bouncy,' and sad ones less so. These differences helped P15 'understand the various emphases the speaker put on words.' At times, though, they felt vibrations and what the text seemed to say were mismatched: 'For example, super intense and strong vibrations when the text seemed bland, unemotional, or matter-of-fact. Or calm, weak vibrations at a particularly emotional moment.' Other participants echoed this, describing how the vibrations could sometimes disrupt their experience. P14 thought some patterns were counterintuitive, noting that 'the worst vibration patterns were very loud and very disruptive to my experience. The speaker would be talking about something personal or serious, and then my wrist is vibrating up the storm.' This suggests that while haptic feedback can help understand emotional content, poorly matched vibrations can lead to a disjointed and distracting experience.

Participants reported that, at its worst, the vibrations caused significant physical discomfort. They particularly disliked the 250 Hz frequency, with P5 describing it as feeling 'like scratching on a blackboard.' P9, who consistently rated 250 Hz poorly regardless of its matching rhythm, said that 'the worst vibrations felt so uncomfortable that I wished the speaker would finish talking. They included too many longer vibrations [LP] or harsh [250 Hz] vibrations at the wrong time.'

The MSP rhythmic pattern, regardless of which frequency it was paired with, also brought negative feedback. Although some participants, like P7, mentioned that it made them feel physical discomfort, the more common complaint touched on it being *distracting*. About MSP pattern, P5 said it 'would vibrate repetitively for one word,' making it very distracting and, thus, harder 'to understand the message in the videos.' P4 found it hard to keep track of words with MSP, a problem they did not experience with other patterns. P16, whose BWS responses had panned both LP and MSP, shared that they 'felt a bunch of vibrations, which kind of overwhelmed me while watching the videos.'

Despite the discomfort or distraction some participants experienced, others found the haptic feedback beneficial in specific contexts. Some participants, for instance, felt it could help them understand an off-camera speaker's emotions. While this was not an issue with the videos used in the study, P9 said that if

a speaker isn't visible, 'this system would help me keep track of their tone and what mood they are in.' P2 echoed this, saying they were 'able to notice the difference in emotion as the person is speaking without visually seeing them.' P5 added: 'with the wrist-worn system, it would be helpful if I could understand whether they are being neutral or emotional when I can't see their face.'

Furthermore, some participants argued that the effectiveness of haptic feedback depends on contextual factors. For example, vibrations could help when speakers communicate with reduced or too subtle facial expressions. P10 thinks 'sometimes hearing people's faces don't really show facial expressions, and I can't tell their emotions.' P3 agreed, saying that intense vibrations would tell if a speaker was 'excited or speaking in a calm manner, which helps deaf people since sometimes hearing people aren't clear with their facial expressions.'

10.4 Discussion of Study 4

In response to RQ4, we found that participants consistently preferred the single short pulse (SSP) haptic pattern over both the long pulse (LP) and multiple short pulses (MSP) patterns. This was clear from the TrueSkill ratings (Figure 10.3) and in some of the comments participants shared.

The picture is less clear when we look at the two evaluated frequencies. While SSP with 75 Hz had a higher TrueSkill than the same pattern with 250 Hz, there is still overlap between the two options' 95% confidence intervals. In terms of the implied probability of choice, this difference means that the SSP 75 Hz pattern has a 62.4% chance of being chosen over its 250 Hz counterpart [173]. For comparison's sake, in a pairing between the top and worst performing patterns – LP 250 Hz and SSP 75 Hz, respectively – the latter would be chosen over the former 89.4% of the time.

While the ratings are close, participants' comments help differentiate the two. Several mentioned that the higher frequency felt physically uncomfortable, comparing it to 'scratching on a blackboard.' This discomfort likely contributed to its lower ratings, and even though the feedback by itself may not be sufficient to entirely discard the high-frequency pattern from future explorations of the haptic design space, here it is enough to justify its exclusion in our second study. Given that, as we will discuss, the second study involved long-form videos, ensuring participant comfort during the test was a key consideration.

While this study focused on participants' subjective preferences, it highlighted both the promise of our proposed haptic-arousal approach – *e.g.*, helping understand the intensity of speakers' emotions – and

some challenges – *e.g.*, potential for distraction. We further explore these themes in Study 5, with a primary focus on identifying an engaging combination of haptic feedback and typographic modulations.

10.5 *Limitations*

Our work in Study 3 suggested modulating either font-size or font-weight to convey arousal. Here, in Study 4, we chose the latter, because it offered better legibility. However, font-size was perceived by some DHH participants as offering a *clearer* depiction of arousal, which could influence how it relates to haptic-feedback also depicting the feature, and thus change the results reported here. This aspect is left as a recommendation for future research.

We conducted the tests in a controlled environment. It is uncertain whether the results would be replicable in different settings, such as varying screen sizes, device types (phones, tvs, etc.), and lighting conditions. This uncertainty extends to the haptic device itself, which was selected based on recommendations from prior literature [64, 196]. Users of these systems may be interested in different configurations, which merits further exploration.

Lastly, the videos and haptic signals were pre-generated. While latency in automatic captioning systems has improved, it is not nil, and it remains to be seen what would be the best strategy to deal with a haptic signal that is out-of-sync with the image of subjects on the screen.

CHAPTER 11

Study 5: Visual–Haptic Affective Captions and Narrative Engagement

Having established the SSP rhythmic pattern combined with the low frequency setting (75 Hz), we set out to answer our follow-up research question and determine whether depicting emotions embedded in speech through haptic feedback and captions with typographic modulations influence narrative engagement for DHH individuals. To do this, we compared four conditions depicting a speaker's emotions through visuals and/or haptics against a neutral baseline with no affective information.

11.1 *Methods*

11.1.1 *Conditions*

To answer RQ5, we defined four conditions regarding the portrayal of arousal:¹ arousal through visuals-only (C_{2V}), haptics-only (C_{3H}), visual *and* haptics (C_{4V+H}), and no arousal depiction (C_{5∅}). We also included a conventional condition that had neither arousal nor valence as a baseline (C_{1B}). Table 11.1 summarizes the five conditions.

¹Because valence isn't our focus here – and Study 3 showed font color conveys it effectively – we encode valence with font color in all conditions.

CONDITION	AROUSAL DEPICTION	VALENCE DEPICTION
C1 _B (<i>baseline</i>)	N /A	N /A
C2 _V	VISUALS	VISUALS
C3 _H	HAPTICS	VISUALS
C4 _{V+H}	VISUALS & HAPTICS	VISUALS
C5 _∅	N /A	VISUALS

Table 11.1: The five conditions presented to participants in this study. The c- abbreviations are used throughout this section. For reference: C1_B are conventional captions (the baseline condition); C2-5 all use font-color to depict valence, with differing approaches for arousal: C2_V uses visuals only (font-weight); C3_H uses haptic-feedback only; C4_{V+H} uses both visuals and haptic-feedback, and C5_∅ uses neither, showing only valence.

11.1.2 *Stimuli*

In selecting videos for Study 5, we followed three basic criteria. These criteria are broadly consistent with those applied in earlier studies (Studies 2–4), but we restate them here because they serve as the foundation for the different choices we had to make in this study, especially regarding arousal variation:

1. The videos should predominantly feature one speaker.²
2. The videos should be short enough to allow all five conditions to be presented within the allotted session time;
3. The videos should tell emotionally charged stories, with a particular emphasis on a variety of arousal levels.

In Study 4, as in studies 2 and 3, the videos used met the first two criteria but generally had unchanging arousal levels. This is understandable, given that the individuals in the SEND dataset were recalling past memories, leading to stories recounted in a calm manner with only occasional bursts of excitement. While for Study 4 we sliced the videos to include these bursts, this approach would not have been suitable for the goals of Study 5. Here, measuring changes in narrative engagement required arousal levels

²This criterion is based on the scrolling-caption-based style from our work in Study 3, which has not yet been adapted or evaluated for multi-speaker settings – a topic outside the scope of this study.

that varied over a longer duration, meaning that a complete narrative arc needed to be established. To achieve this, we selected fictional videos with diverse arousal levels that could also tell a full story.

We searched both general (*e.g.*, YouTube, Vimeo) and short film-dedicated platforms (*e.g.*, ShortVerse, Short of the Week).³ From an initial pool of 26 titles, we processed 13 through the affective captioning pipeline (described in subsection 9.1). These were evaluated by a second rater and me for narrative coherence and consistency between perceived and synthetically inferred emotions. Ultimately, five videos were selected, as listed in Table 11.2. Each of these five videos was prepared in all five conditions, giving us 25 combinations to counterbalance each story’s inherent effects on narrative engagement.

NAME	SOURCE	ORIGINAL SOURCE
<i>Sheriff Hassan’s Monologue</i>	<i>Midnight Mass</i> , episode 6, season 1	youtu.be/0lhpqJso4tM
<i>Sally’s Monologue</i>	<i>Barry</i> , episode 7, season 2	youtu.be/qw62N4v8Cwo
<i>The Arrival</i>	Short film by Daniel Montanarini	vimeo.com/166075559
Scene from <i>Damage</i>	Short film by Matt Porter	vimeo.com/325243238
Scene from <i>The Human Voice</i>	Short film by Pedro Almodóvar	DVD copy

Table 11.2: The five videos used in Study 5.

11.1.3 Narrative Engagement

A challenge in conceptualizing *effectiveness* in affective captions, whether coupled with haptic feedback or not, comes from defining what it is that they allow their users to do better when compared to traditional captions. Previous, we – along with other researchers [85, 99] – have relied on self-reported measures of usefulness (Study 2) and objective assessments of perceived valence and arousal levels (Study 3). These methods provide valuable insights, but they also have limitations. Self-reported measures may be susceptible to novelty bias [198], and assessing the interpretation of valence and arousal levels does not necessarily indicate whether users’ engagement with the content is actually affected by having access to this additional information. Because of these points, in Study 4 we propose using a different metric to capture the effects of affective captions: *narrative engagement*.

Narrative engagement, as a measure, captures changes in cognitive processes in individuals as they attempt to make sense of a story [35]. It builds upon established constructs such as *spatial presence*

³shortverse.com and shortoftheweek.com

(the sensation of being physically present and able to act within the depicted environment [108, 203]), *identification* (experiencing events portrayed in the narrative as if they were happening to oneself [43]), *flow / transportation* (becoming deeply absorbed in the narrative to the extent of losing self-awareness and awareness of surrounding events [34]), etc.

Although narrative engagement instruments are typically used to explore the phenomenological aspects of engagement with fictional stories, the cognitive processes they model are not limited by the distinction between fiction and non-fiction [162].⁴ As Gilbert suggests, human perception accepts mental representations as true, and disbelief requires additional cognitive steps [74]. This implies that although we used fictional videos in our experiment, one can reasonably assume that similar effects could be seen with a similar setup used in more general contexts.

Narrative engagement has been widely used in media studies and HCI research to compare how different platforms influence audiences' experiences. For example, studies have compared the experience of watching a 360° video using virtual reality headsets versus smartphones [25] (finding no significant differences), evaluating game narratives with low versus high fidelity graphics [31] (also finding no significant differences), and comparing automatic versus professionally authored closed captions for YouTube videos [98] (again, finding no significant differences).

While this diverse set of comparisons did not yield measurable significant differences, it does not undermine the utility of the narrative engagement instrument. Instead, it highlights the robustness of the underlying processes it measures, which seem to transcend variations in media fidelity. This is intuitive for anyone who has been absorbed in a book – a notably low-fidelity medium that is nonetheless capable of eliciting deep immersion. However, there are cases where manipulations do produce measurable effects: for example, Sukalla et al. experimentally varied cohesion and emotional content in medical drama clips and found significant differences across subscales, with lower cohesion reducing narrative understanding and presence, and higher emotional content increasing emotional engagement and attentional focus [175].

⁴A difference between the two, argues Busselle and Bilandzic, is that we use different schemas – *i.e.*, the stereotypes and tropes we bring in as predetermined expectations about how events will unfold – to process fiction and real-life [34].

11.1.4 *Experimental Design*

Study 5 employed a single-factor, univariate within-subjects design to evaluate the effects of haptics, typographic modulations, and their combination on narrative engagement. Each participant experienced all five conditions, randomly applied to each one of the five videos to account for potential confounds arising from the inherent narrative engagement of each video or potential asymmetric transfer effects. By counterbalancing the order of presentation, we aimed to mitigate the influence of specific video content on the participants' engagement scores and isolate the effects of the captioning conditions.

11.1.5 *Experimental Procedure*

Like with the previous study, this was conducted in person. A research assistant fluent in ASL met with participants and, after the introduction and consent procedure, helped them attach and calibrate the haptic device. After this, participants went through the five videos, presented in counterbalanced order and conditions, responding after each one the 12-item narrative engagement instrument. The questions used, grouped by their four sub-scales, are presented in Appendix B. Each question was presented as a Likert-type item, allowing participants to indicate their level of agreement on a scale ranging from 1 to 7. In the analysis phase, some items were reverse coded so that higher scores consistently reflected greater narrative engagement [35].

After finishing this section, they were presented with a screen, shown in Figure 11.1, that described each of the five conditions and which video they were applied to, and asked three open-ended questions about each condition, namely, *Did you think this caption style worked well with this particular video? Why, or why not?*, *Did you like this caption style? Why, or why not?*, and *In what genres of video or viewing situation do you think this caption style would work well? E.g. 'Watching a sci-fi movie at the cinema.'* To analyze these answers, we used a *inductive thematic analysis* method, where one of the authors engaged with the data, allowing patterns and central ideas to emerge from participants' responses [33]. These were then discussed with other authors and refined.

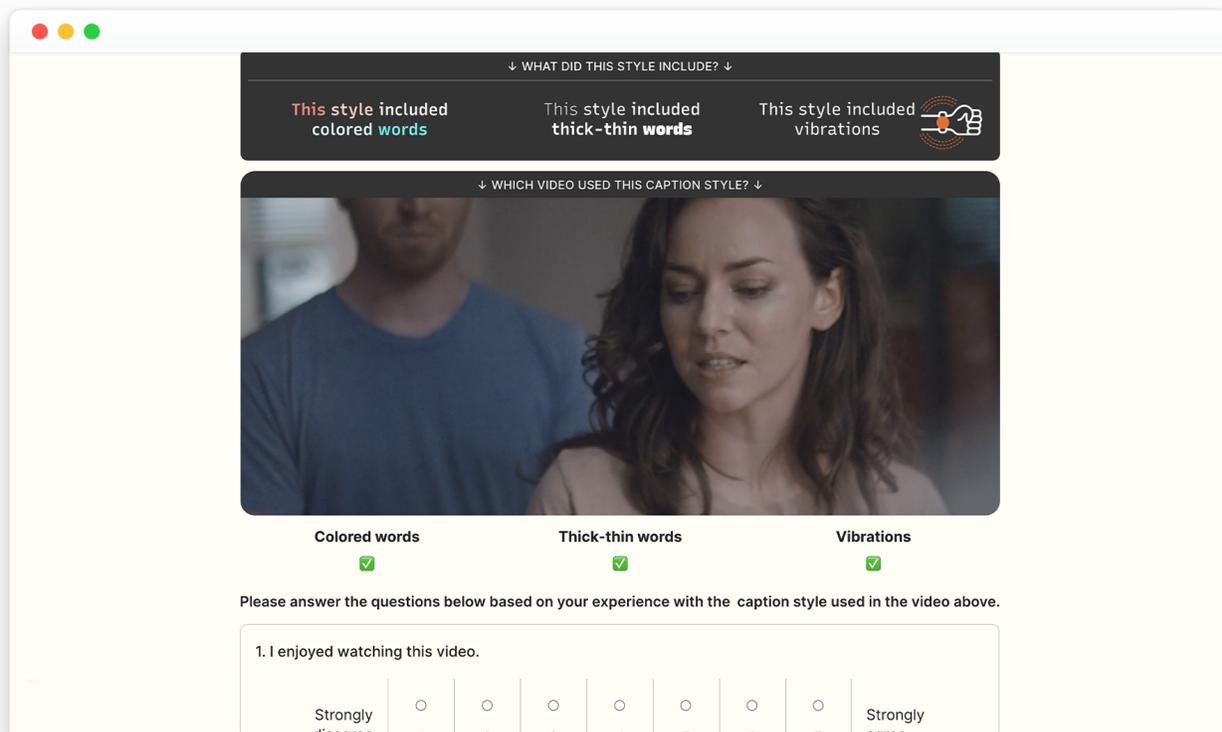


Figure 11.1: Screenshot of the page where participants gave feedback about each condition. The image captured from the video was used as a mnemonic device for each caption condition, together with the illustrations and short descriptions. In this example, we see $C_4 V+H$ (here labeled as ‘caption style 1’), which includes visuals and haptics for arousal, and visuals for valence.

CONDITION	NARRATIVE ENGAGEMENT				TOTAL SCORE
	NARRATIVE UNDERSTANDING	ATTENTIONAL FOCUS	NARRATIVE PRESENCE	EMOTIONAL ENGAGEMENT	
C1 _B	15	14	12	15	51
C2 _V	16	14	11	13	55
C3 _H	16	16	11	14	54
C4 _{V+H}	18	15	13	16	62
C5 _∅	17	15	13	14	60

Table 11.3: Median raw scores for each of the four sub-scales and median total Narrative Engagement score. See Figure 11.2 for distribution of scores for each condition. Note that each sub-scale ranges from 3 to 21, and the total scores range from 7 to 84.

11.2 Findings from Study 5

Recruitment, compensation, and inclusion criteria matched those of Study 4. We initially had a total of 31 participants, although one was removed from the data due to equipment malfunction during testing, with three others excluded after test duration logs indicated that they had not watched all of the stimuli videos in full. Among the remaining 27 participants, 15 identified as female and 12 as male, with 20 identifying as d/Deaf and 7 as Hard-of-Hearing. Their mean age was 24.7 ($\sigma = 7.6$).

Throughout this section we follow the condition-abbreviation scheme presented in Table 11.1. Where needed, participant quotes were edited for clarity and conciseness.

11.2.1 Narrative Engagement

Scores were initially calculated summing the 12 raw Likert-scale items for each condition for each participant. The median values for overall Narrative Engagement scores for each one of the five conditions, as well as median values for the sum of each of its four sub-scales, are presented in Table 11.3.

The distribution of scores is shown in Figure 11.2. Given that this is a within-subjects study with non-parametric data,⁵ we used the Friedman test to compare the raw answers – *i.e.*, the individual 12 Likert-

⁵While there are examples both ways, we align ourselves with authors who have treated narrative engagement data as non-parametric, *e.g.*, [77, 78, 206].

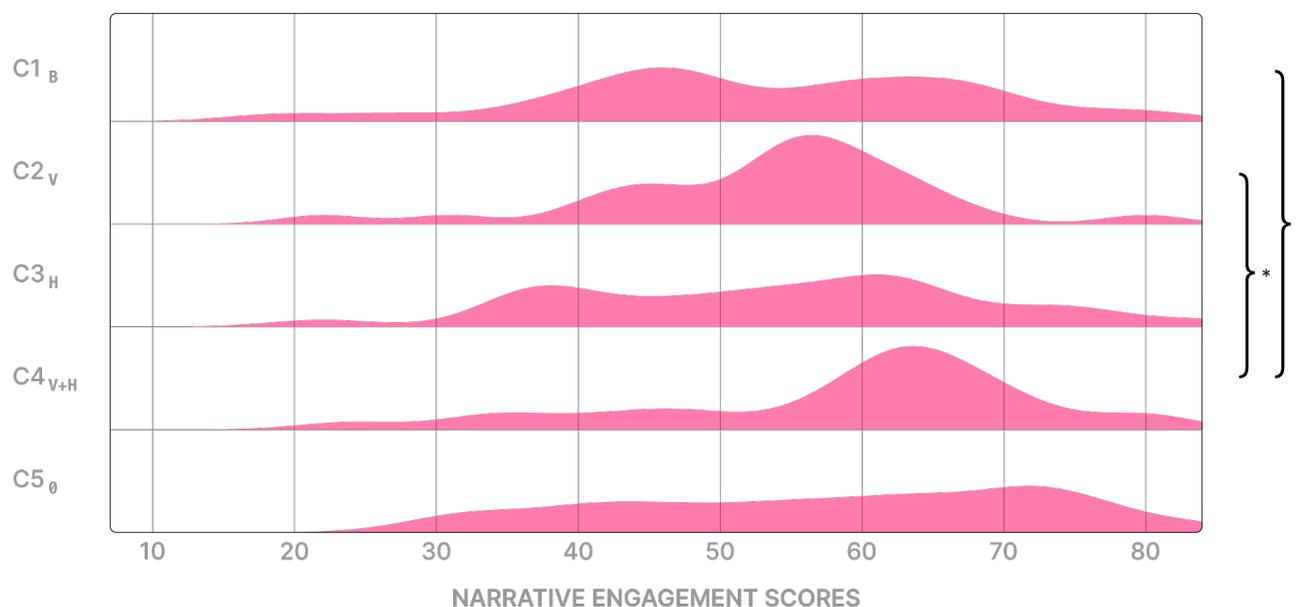


Figure 11.2: Ridge plot of Narrative Engagement scores across conditions, ranging from 7 to 84. Each ridge represents a condition, with its height indicating the density of scores. Significant pairwise comparisons ($p < 0.05$) between $C4_{V+H}$ and $C2_V$, and $C4_{V+H}$ and $C1_B$, are highlighted by the curly brackets.

item narrative engagement scores answered for each condition – with Dunn-Bonferroni post-hoc pairwise comparisons for significant results.

The Friedman test indicated statistically significant differences in the narrative engagement scores across the different conditions ($\chi^2(4) = 32.5, p < 0.001$). Post-hoc comparisons showed significant differences between $C1_B$ and $C4_{V+H}$ ($p = 0.01$) and between $C2_V$ and $C4_{V+H}$ ($p = 0.02$). The Friedman test yielded statistically significant differences for the *Narrative Understanding* ($\chi^2(4) = 12.0, p = 0.02$) and *Narrative Presence* ($\chi^2(4) = 18.4, p = 0.001$), but with no significant differences in the post-hoc tests.

11.2.2 Open-Ended Data

Much like with their answers to the *narrative engagement* questionnaires, opinions on the five different captioning conditions were widely spread, ranging from participants sharing how their feelings were shaped by the haptic and visual cues as they watched the videos, to some that questioned the premise of having an external source of interpretation for characters' emotions, to many others who commented

on ways they felt the different approaches could be made better. In this section we present three themes that organize these thoughts.

11.2.2.1 THEME 1 *Establishing an Emotional Connection*

For some participants, adding haptic feedback improved how understandable the videos were and how connected they felt to them. Discussing C_{4V+H} , P1 said that ‘the haptic feedback does make a difference. I feel I can understand the movie without context. I was able to pick up the story.’ Haptics, they added, made it ‘feel more exciting and connected to their world.’

Visuals also helped. P1 said that they ‘would work great in a scene where there were multiple people talking with differing emotions and tones. The colored words (C_{2-5}) and emboldened fonts (C_{2V} and C_{4V+H}) could make clear what is going on in the scene and the emotions being shown on screen.’ For P10, affective captions bridged information gaps left by other communication channels. They complimented C_{2V} because of how it helped them ‘understand the character’s tone even though they were displaying almost no facial expressions in the scene.’ For P25, C_{2V} allowed them to see the character’s emotions, which helped them ‘get more connected to the speaker’s voice.’ P12 thought C_{2V} showed character’s ‘emotions and tone of voice,’ saying that they were ‘not used to this new caption, [but] it helps me understand better.’

Commenting on the baseline condition (C_{1B}), some participants shared that experiencing the other modalities made conventional captions feel lacking. P10 thought that going back to C_{1B} ‘was unusual because every other video prior had some feature to make me feel involved, and then this one was just black and white, and I had to rely on the speaker’s facial expressions... I think it’s still nice, but once I see the colors and thickness, those styles were more engaging than this one.’ This reliance on other visual cues becomes apparent. P14 said that with C_{1B} they were ‘unsure what the emotions are,’ needing ‘to rely on context that is surrounding the character to identify (guess) their emotions.’

Not all participants found haptics helpful. For some, the constant vibrations were a distraction, which drew them away from the scenes. P25, commenting on C_{4V+H} , said they ‘felt the haptics were too overwhelming, distracting me from connecting with the speakers,’ going so far as to say that ‘haptics ruined it.’ P13, explaining why they preferred C_{2V} over the conditions with haptic feedback, said that ‘vibrations have a purpose, but I feel that they distracted me from the story. In this particular scene, they would have taken away instead of adding to it.’

Fast speakers can make this distraction worse. Commenting on C_{3H}'s use in *Sally's Monologue*—a frantic stream-of-consciousness monologue—P6 said that ‘in this kind of video, the speaker talks really fast,’ adding that in that case ‘vibration is a distraction to me, and it is difficult to follow captions when you can't concentrate.’

Improvements to feelings of empathy and spatial presence facilitated by affective captions were also discussed. Adding to their impressions of C_{3H}, P7 said that they ‘could feel what the speaker was feeling through colored words, and I understood their pain as if it was my own, so it worked very well.’

Some participants described feelings of spatial presence. P10 felt C_{4V+H} made them ‘see the character's thoughts vividly,’ making it their ‘favorite combination out of all the five videos. The combination of the colors, thickness and thinness, along with the vibration, made me feel like I was truly there in the scene.’

11.2.2.2 THEME 2 *Affect as Information*

There was pushback from some participants about how well affective captions could work. For one, there was uncertainty about how precise such a system could be. P22 talks about how, while the visual cues in C_{2V} were clear, they were not sure whether they were necessarily accurate, or whether they were able to ‘100% convey the speaker's original emotion,’ which led them to ‘disassociate slightly from them.’

This is further complicated by how complex emotions can fall outside, or become ambiguous, under the circumplex model of emotion. Again P22: ‘There was also a hint of sarcasm somewhere and I'm not sure if it was really captured with the subjective colorings, because emotions are subjective.’ P21 suggested that ‘the caption style could include a few different types of emotions. Something to show the character's depression, disgust towards certain things, patience with his life, not just sadness or rage.’ This lack of nuance was also seen by P8 who, commenting on whether C_{2V} was able to match one of the videos, said that while the style allowed them to ‘understand the environment of the video,’ the visual cues ‘seemed to be out of tune. Is the speaker being sarcastic or genuine?’

In some of these cases, this mistrust seemed to be related to a mismatch between what the visual cues and/or vibrations were telling and what viewers were getting from other cues in the videos. Commenting on C_{3H}, P21 tells that they were ‘a little lost because the character almost didn't show emotions, even though the caption style showed their feelings.’ P7, discussing C_{2V}, talks about how ‘red text tells us the speaker is feeling angry or somewhat frustrated, and then when the text turns to bold it made me

think that the speaker is shouting, but the speaker is actually thinking to themselves in the video, so that connection between the text and the speaker isn't there.'

Some of the resistance stemmed from the intended purpose of affective captions, namely, to provide an external interpretation of emotions as conveyed by a speaker's tone of voice. P22 offers that 'emotion is subjective, and it is up to the viewer/listener to interpret it, so I'm not sure if it is necessary for a captioning user interface to determine that.'

For some, the need for visual or haptic affective cues depends on whether a speaker's emotions are otherwise clear. In the *Hassan* video, for instance, P15 felt the added affective cues were not needed because the speaker was 'able to express their emotions in a sincere and clear way to those watching.' The neutral-looking C1_B would be better, they added, 'for those who are skilled at acting and conveying emotions not just by tone of voice, but by facial expressions as well.'

As a counterpoint, many participants embraced the information that captions added to the scenes. P13, on C4_{V+H}, says that 'the three aspects serve as a great supplement to the story, as they gave me a good idea of just what the speaker felt.' P11 said they loved the idea behind C4_{V+H}, since 'films show purposeful and powerful emotions, and all of us, especially people who read closed-captions, would like to be part of it.'

For some participants the value added by affective captioning approaches, in particular for conditions that had haptics, seemed to be not necessarily because of the exact emotions they conveyed, but rather because of how they highlighted *shifts in moods*. P14, commenting on C3_H, offered that he felt it worked 'because I was able to see the start and finish of the emotions the character plays.' Echoing this sentiment, P23 mentions that C4_{V+H} 'worked well because it allowed me to understand the shifts in the speaker's mood and attitude over the span of the video.'

11.2.2.3 THEME 3 *Contextual Considerations in Affective Captions*

Some feedback focused on how effective the visual and haptic parameters implemented in C2_V, C3_H, C4_{V+H}, and C5_∅ were. While part of this was included in the two previous themes as it relates to their own overall discussions, a few comments had a narrower scope, focusing on a deconstruction of the design underpinnings of affective captions and how they work (or don't) under different situations.

As previously noted, many participants found the use of vibrations distracting. This was also true for font-color. For instance, P26 felt that C5_Ø had ‘too many colors in one sentence, making it easily distracting.’ For P18, the use of color worked and was able to influence how they perceived emotions, but it also made text ‘hard to read while I was thinking about other things.’ P14 was even harsher: ‘I don’t think this caption style worked because I couldn’t figure out what the colors represent. It just felt like an update to the current captioning style, but nothing really changed.’

Some complaints focused on the colors used for positive and neutral tones. P9, for instance, complained that ‘it is kind of hard to see the difference between blue and white,’ as did P17, who disliked C2_v because it was ‘hard to recognize the blue or white font, making it hard for me to identify happy and neutral tones.’ For P3, ‘caption color was not vibrant, so it was hard to decipher.’ P1 found this particularly tricky with lighter shades of red and blue, stating, ‘I wasn’t sure if some words were neutrally white after long reading. It was almost like they blended together. I think it needs some adjustments to color tones and fonts.’

The use of font-weight also had its discontents. P24 thought that, in C2_v, the ‘color changes and bolding of captions hurt my eyes.’ Many participants complained that the more extreme font-weights used in the captions did not work. P26 thought ‘the font is too thick to recognize,’ while P20 said that it ‘made everything feel blurred.’ At times this was caused by the combined effect of having changes to weight and color. P5 thought ‘bold is too much when the caption is also colored.’ P1 echoed this: ‘the font with the bold felt almost hard to read along with the color.’

Some participants were not against the idea of haptics, but felt it could be used only for important words, or even for non-speech sounds. P9 suggested that ‘in horror movies, it could include only the screaming. Suspense, but not the words.’ On this, P23 added: ‘Maybe it would be a good idea to limit the vibrations to only the emotional climaxes in movies. Having vibrations on throughout the whole movie would probably be distracting and annoying.’ P25 went further, saying that in sci-fi or scary movies it could be used ‘so we can feel background noises, scary music, etc.’

11.3 *Discussion of Study 5*

11.3.1 *Using Haptic Patterns to Convey Arousal*

We saw that the fourth condition (C_{4V+H}) stood out as the most effective. This condition combined the winning haptic pattern from Study 4 with visual modulations of font-weight for arousal and font-color for valence, as inspired by previous research. This haptics-visuals integrated approach significantly outperformed a visuals-only affective caption style (C_{2V}), which was designed to mirror previously discussed affective captioning models, *e.g.*, [53, 85, 99]. Interestingly, our findings suggest that a combination of both haptics *and* visuals creates an experience that, for our 27 DHH participants, resulted in higher levels of narrative engagement with audio-visual content. Thus, in answering RQ5, we recommend a *combined* approach to haptics and visual modulations to depict a speaker's arousal levels.

Furthermore, we found that the condition combining haptics and visuals also promoted significantly higher narrative engagement scores when compared to the baseline condition (C_{1B}), *i.e.*, conventional, non-styled captions. In other words, the C_{4V+H} option was more engaging than both the conventional captions in everyday use and the recommended option from prior work on affective captions (C_{2V}).

11.3.2 *Consideration of Users' Experience with Affective Captions that Employ Haptic Feedback*

Despite quantitative findings showing that the combination of haptics and visuals led to a significant improvement in narrative engagement, our participant' feedback revealed individual variability in their subjective experiences. While some participants found the vibrations to be a valuable addition that enhanced their connection to the videos, increased empathy, and created a sense of spatial presence,⁶ others experienced the constant buzzing as a distraction that pulled their attention away from the content, disrupting instead of improving their overall viewing experience. This echoes a finding that echoes Wang et al. [196], who previously combined haptics and captions to aid with speaker identification. While this diversity in users' experiences could simply be a byproduct of the inherent diversity within the DHH population in general [172], the tension between enhancement and distraction also aligns with broader challenges in designing multimodal captioning systems [27], particularly when considering the cumulative effects of such features over longer durations.

⁶Spatial presence in this context refers to the user's perceived sense of physical existence within a digital environment, where the technology facilitates a feeling of being 'there' in the virtual space, contributing to a more immersive experience.

Our study's design, which featured multiple conditions and extensive surveys, did not accommodate long-length videos. There are reasons to believe that the effects we measured could be even higher in such settings. For one, longer videos are correlated with higher Narrative Engagement scores [93], so the differences between conditions we saw could potentially accumulate over time. This could compound with how decoding affective captions is subject to learning effects, as we saw in Study 3, or with how certain haptic stimuli may become more favored through repeated exposures [92]. However, it remains to be seen whether the distraction and annoyance that some participants experienced would persist over time. While these issues could plausibly subside given *sensory adaptation*, *i.e.*, the phenomenon where sensitivity to a haptic stimulus diminishes after prolonged exposure [17, 148], they might also continue or even intensify depending on individual differences. This underscores the need to study whether sensory adaptation lessens distraction over time or if prolonged exposure increases annoyance, both of which could affect narrative engagement.

Related to this point, future work could also look into thresholding approaches to mitigate the negative aspects some participants experienced, such as distraction and annoyance. If haptic vibration were to occur only when some relevance threshold was crossed – *e.g.*, only vibrate words that are significantly more intense or calm than the average – then perhaps distraction can be minimized. Adjusting intensity dynamically or via user-based personalization could also help, although additional studies would be needed to further explore this.

Focusing on the four sub-scales, we see that while the Friedman test revealed significant differences for the *Narrative Understanding* and *Narrative Presence* sub-scales, no significant differences were found in the post-hoc analysis across the five conditions. This suggests that while *Narrative Engagement* can give a comprehensive measure of engagement with the audio-visual content, it may also be too blunt a measure for an in-depth exploration of its four sub-scales independently. For such purposes, more targeted instruments might be preferable – a recommendation for future research.

Quantitative data and participant feedback indicate that haptics were especially effective when paired with visual arousal cues, suggesting *intermodal integration* – stimulation in one channel can enhance or alter perception in another [26]. We echo Kushalnagar et al. [105], who found that visual-tactile captions for non-speech information outperformed tactile-only ones. This effect may explain the non-significant advantage⁷ of C4_{V+H} over C3_H, where haptics and visuals outperformed haptics alone for conveying

⁷Although these differences were not statistically significant, we offer speculative commentary for future research.

arousal. The higher performance of C_{4V+H} over C_{3H} further suggests that haptics alone may not provide sufficient perceptual salience for arousal, underscoring the importance of intermodal cues.

Alternatively, the non-significant patterns observed in the improvements for $C_{5\emptyset}$ – which had *no* depiction of arousal – over both C_{2V} and C_{3H} could suggest that, if arousal is not strongly reinforced by both visuals and haptics, it might be better to omit it altogether. This could be related to how arousal has been shown to be perceived as if of lesser importance than valence [62]. While future work should explore this hypothesis further, the relative performance of the conditions also point to a novel direction for research: if intermodal integration in the C_{4V+H} condition is effectively facilitating the communication of arousal as a speech dimension, could similar strategies enhance the depiction of valence through haptic signals? For example, would modulating the frequency in tandem to valence, alongside the amplitude changes that convey arousal, reinforce the font-color modulations, increasing the perceptual salience of valence?

For some participants, the haptic feedback acted not merely as a synthetic signal but as a direct analog of the speech signal itself. This suggests that, to them, haptics was perceived as a form of *sensory substitution*, *i.e.*, they understood the vibrations as if representing the actual speech sounds, instead of an artificial signal that is related, but not equal, to speech. This approach aligns with Wang et al.'s method for haptic captions [196]. While the extent of this perspective among DHH users of affective captions remains to be fully explored, it presents a promising avenue for future research: could the actual sound signal, *i.e.*, its amplitude envelope and frequencies, be an additional dimension in the haptic signal? Should this dimension replace the synthetic arousal signal, or be integrated into it?

11.3.3 *Fine-tuning Color and Font-weight Style Dynamics in Affective Captions*

Participants' feedback on affective captioning styles also relates to design guidelines already established, *e.g.*, those from Study 3 and Hassan et al. [85]. Questions arise on how clear the colors used are, but also how much they should leave open to interpretation; in terms of font-weight, guidance is needed to answer: how much is too much? It was positive to see that the legibility of the captions we used did not emerge as a major concern, which is an improvement over similar past studies that were plagued with these issues, *e.g.*, studies 2, 3, and Kim et al. [99]. Still, there were cases where the font-color and weight modulations did not work well.

The color palette recommended by Hassan et al. [85] appeared ambiguous for near-neutral words. This need not be necessarily seen as a defect. Given how affective information can be thought of as context-dependent [29, 87], some researchers have advocated that design solutions are made to be purposefully ambiguous and open to interpretation [72]. In fact, some participants appreciated the color scheme's flexibility for personal interpretation. However, complaints could also reflect disagreement with the chosen colors. Future work could explore alternative palettes that better balance clarity with openness to contextually-based interpretations.

While our findings from Study 3 suggest the use of changes to font-weight to depict arousal, further work would still be needed to define specific guidelines for how these should be implemented. While minor changes in weight may not significantly affect legibility [145], in our implementation, words with very low or high arousal were shown with extreme weight changes, which some participants felt was too much. Future work should establish clear thresholds for designers. Additionally, we observed negative effects from certain combinations of visual modulations, as P5 noted the overwhelming impact of bold fonts used alongside colored words, meaning color should be included as a confound in these studies.

These findings serve to both affirm the current design guidelines on how to use typographic cues to convey emotional content in text and to suggest that more precise recommendations are still needed to optimize and ruggedize the application of affective captions. This underscores the importance of iterative testing and refinement in their design as more and more scenarios and use-cases are explored. Alternatively, the variance in opinions could be seen as a case made for offering personalization options for the visual parameters. In this, they would echo May et al. [125], who has suggested that a one-size-fits-all approach for non-speech information accessibility may not be sufficient given how DHH expectations and preferences vary.

11.4 *Limitations*

As a construct, *Narrative Engagement* is reflective of users' experience, and can be tied to improvements to the perceived *dullness* in traditional captions, as we saw in Study 1. It does not capture other measures, such as comprehension, memory, task performance, etc. Some of these are related to how captions are also perceived as *ambiguous*, and as such is an important consideration for future studies.

The materials and setting also bound generalizability. Stimuli were single-speaker, scripted clips of short to medium duration, viewed in a controlled, low-distraction environment. Effects may differ with multi-speaker dialogue, overlapping speech, live/unscripted content, longer formats (*e.g.*, feature-length), different genres, languages, or typical at-home contexts with competing distractions.

Our implementation choices further constrain interpretation. Haptics were delivered on the wrist via a watch-like device using a single rhythmic pattern at 75 Hz; other body locations, devices, or parameterizations (*e.g.*, different carrier frequencies, envelopes, or adaptive intensities) may yield different results. Visual encodings used a specific color palette for valence and font-weight changes for arousal; legibility and discriminability can vary with display characteristics, ambient lighting, and individual differences (including color-vision deficiencies). Several participants reported distraction with stronger weights and dense color changes, suggesting thresholding or rate-limiting may be necessary in longer viewing.

CHAPTER 12

Study 6: Haptics-decoding Performance¹

Study 4 identified a haptic pattern that, through its combination of frequency, rhythm, and varying intensity, was judged by participants to be a good encoding of speech arousal levels. When used alongside typographic cues that visually encoded emotions, this pattern improved viewers' engagement with audiovisual content. This suggested that affective captions could meaningfully enhance the experience of DHH viewers by addressing one key limitation of conventional captions, as we had seen in Study 1: they are often *dull*.

Yet, while engagement captures how drawn in or emotionally attuned viewers feel, it does not necessarily tell us how well they are able to *decode* or make sense of the conveyed information. In Study 1, participants highlighted that the *ambiguity* of conventional captions is a key pain point. If the goal of the cues we add through visuals and haptics is not only to improve engagement but also to help disambiguate speech and speaker intent – and it is – then understanding whether viewers can make sense of them is an important piece of the puzzle. In Study 3, through the EmojiGrid experiment, we saw that participants could successfully interpret information encoded through visual changes to font color, weight, and size. However, after Study 4, it remained unclear whether the same could be said of the haptic signal. In other words, how clearly do participants actually understand the arousal values conveyed through haptics?

Addressing this question not only complements our findings on engagement but also deepens our understanding of what makes affective captions effective. If captions aim to improve how *understandable* speech – in all its complexity – can be, then whatever information is conveyed through haptics must also be understandable. Moreover, this knowledge could guide future efforts to use haptics to represent

¹This study was jointly conducted by me, Stephanie Patterson, graduate student at RIT, and my co-advisors, Dr. Roshan L. Peiris and Dr. Matt Huenerfauth. I led the study design, stimuli creation, data analysis, and writing of the paper, which is currently being prepared for submission.

other kinds of information, whether for speech accessibility [196] or in broader multimodal contexts [124], thus expanding the expressive potential of the medium.

As such, while Study 5 examined how multimodal captions affect viewers' emotional engagement, Study 6 turns to the question of *decoding*: viewers' ability to extract or recognize the intended affective meaning encoded in the cues. Whereas engagement concerns how immersed or emotionally connected viewers feel, decoding concerns interpretability – the clarity and accuracy with which information is understood. This distinction is key, as improving interpretability directly addresses captions' ambiguity, one of the main issues identified in Study 1.

To guide our investigation, and considering our multimodal setup, we propose the following research question:

RQ6 How well do modulations of haptics and typography map to how DHH participants perceive a speaker's arousal levels?

To answer it, we designed a study where participants would be exposed to short video clips where the speaker's voice would present itself in arousal levels from across the spectrum – ranging from very low to very high. In these clips, these levels would be conveyed either through haptic feedback, typographic modulations, both, or neither, allowing us to measure how effective each modality is in conveying the affective dimension.

12.1 *Methods*

Our set-up for this study was similar to that of Studies 4 and 5: participants wore the haptic device on their non-dominant arm while watching short video clips on a computer screen. The videos' original soundtrack was available and played through an external speaker. Unlike our previous studies, participants were allowed to wear the device on either side of the wrist. This flexibility was acceptable here because we were not comparing signals of different frequencies and therefore did not need to account for differences in the skin's frequency response across the ventral (palmar) and dorsal (back-of-hand) surfaces. Figure 12.1 shows this set-up.

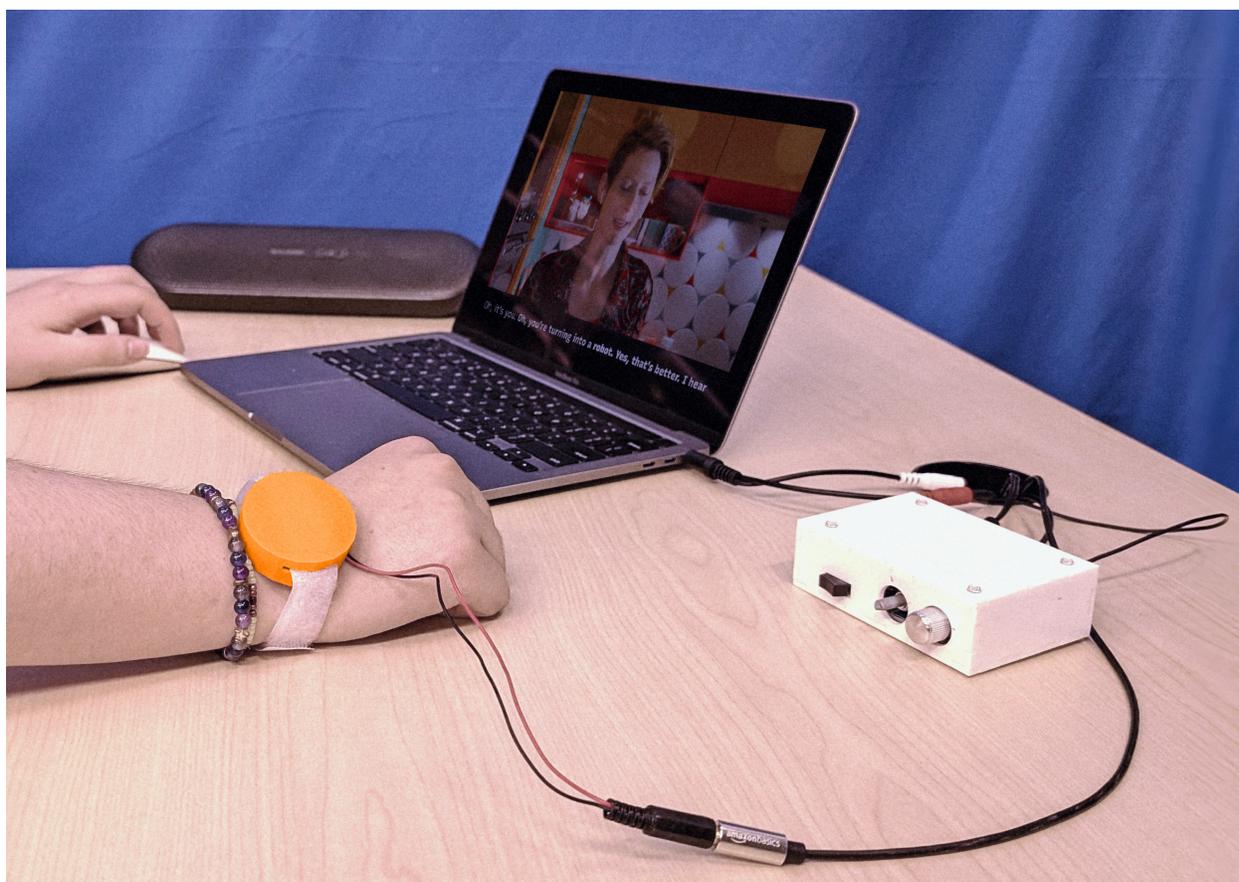


Figure 12.1: Picture of the set-up for Study 6.

12.1.1 *Conditions*

Participants completed an arousal-identification task under four within-subject conditions, balanced across trials: (1) *No visuals, no haptics* (plain captions, no haptic signal); (2) *Visuals only* (caption weight modulated by arousal, no haptics); (3) *Haptics only* (wrist-worn vibrotactile feedback modulated by arousal, unmodified, plain captions); and (4) *Visuals & Haptics* (both channels modulated).

12.1.2 *Stimuli*

We assembled a pool of captioned clips that included a TED talk, a talk-show host giving a monologue, an interview with an Iranian filmmaker, scenes from movies (short, medium, and feature length), a recorded Zoom meeting, among others. The content is challenging to summarize, given that the clips were very short (between 2 and 12 seconds) and that they came from 22 distinct sources, but as a sampler they included a tense phone call with an ex-lover, a woman singing while driving, a man reflecting on his life, a eulogy for one's deceased mother, a cheerful discussion about fashion choices, a tense negotiation for the sale of a bass amplifier, etc. The excerpt selection is discussed below, but the videos themselves were chosen to ensure diversity in speech style (natural or acted), video genre, and speaker gender and age. An example of such an excerpt, together with the test platform, is shown in Figure 12.4.

To determine which excerpts to include, our rationale followed two main principles:

1. We sought to capture segments representing different points along the arousal spectrum – from calm and subdued to highly energetic – so that the full set of videos would include examples covering the entire dynamic range of the affective dimension.
2. Because participants would be asked to estimate a single average arousal level per clip, we prioritized segments with low internal variability in arousal over time. This was crucial for interpretability: if a segment contained strong fluctuations, participants might anchor their ratings on different portions of the clip.

A practical consequence of these two principles was that we favored shorter clips, from which it was easier to identify candidates that both spanned the arousal spectrum and maintained low within-clip variability. This was also well suited to the temporal nature of haptics: because the vibration cue is

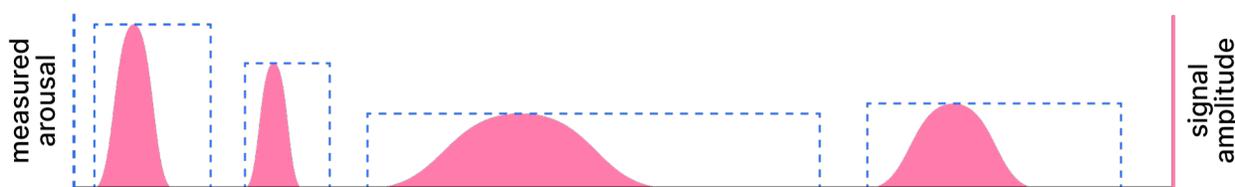


Figure 12.2: Example of the Single Short Pulse (SSP) haptic pattern, originally explored in Study 4.

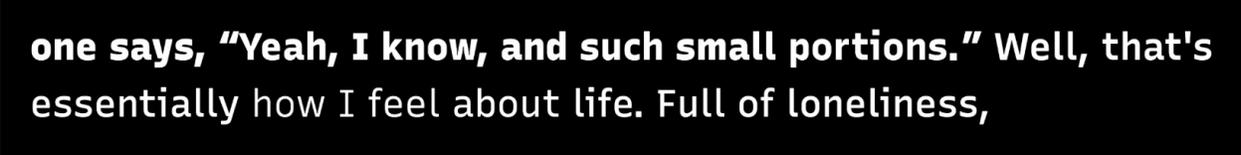
experienced sequentially over time – unlike a caption line that can be apprehended at a glance – short, locally stationary segments helped give focus to the moment-to-moment haptic impression. This meant that, at the lower end, videos were as short as two seconds, a deliberate choice aimed at preserving low within-clip variability in arousal and to allow participants to form a single, coherent impression of intensity.

The dialogue from the videos was transcribed as WebVTT with word-level timestamps using Whisper [151] and its timestamped extension [117]. Each token was annotated with a dimensional arousal score $a_i \in [0, 1]$ by Wagner et al. [190], which we encode in the caption classes (e.g., `v_arou_0p565`) following a coding approach I describe in de Lacerda Pataca [49] and have used throughout this dissertation.

Using token durations, we slid a 20-token window with a hop of one token to create candidate segments. Windows had to be temporally continuous (no gaps > 5 s) and contain at least six valid tokens. For each valid window we computed a duration-weighted mean arousal and its duration-weighted spread. We partitioned the arousal range into six target levels $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and, within each source video, selected the window nearest each target (tolerance ± 0.20) to ensure broad coverage with enough candidates per bin while avoiding excessive overlap, breaking ties by lower within-window variability. This yielded a bank of comparable, low-variance clips across the arousal spectrum.

12.1.2.1 Haptic Mapping

Vibration amplitude followed the inferred arousal at the *word* granularity. Following Study 4’s recommended SSP pattern, shown in Figure 12.2, each word was rendered as a 75 Hz sine pulse lasting two-thirds of the word’s duration followed by a one-third inter-word pause; amplitude scaled linearly with the token’s arousal value $a_i \in [0, 1]$. The signal chain was audio interface \rightarrow Techtile Toolkit amplifier \rightarrow Acouve Vp2 actuator enclosed on the non-dominant wrist.



one says, "Yeah, I know, and such small portions." Well, that's essentially how I feel about life. Full of loneliness,

Figure 12.3: Example of the font-weight modulation, originally explored in Study 3.

12.1.2.2 *Visual Mapping*

We followed the font-weight modulation approach described in Study 3, with an example shown in Figure 12.3. Captions used the Recursive typeface [140] with font-weight mapped linearly to arousal (low $a_i \rightarrow 300$; high $a_i \rightarrow 1000$). No other caption attributes were altered.

12.1.3 *Experimental Design*

The arousal-identification task used a 2×2 within-subjects design (Visuals: on/off; Haptics: on/off). In eight rounds, participants viewed sets of short clips spanning the six target arousal levels, *i.e.*, 48 total measurements, with clip-level assignments counterbalanced across the four conditions.

On each trial, participants rated the speaker's perceived arousal using a continuous slider anchored at *calm/subdued* and *intense/animated*, with its numeric value (0–10) recorded as the trial rating r . The slider interface is visible in Figure 12.4, alongside the captioned video. Clips could be replayed before a rating was submitted.

To mitigate order and assignment effects, stimuli sequencing and condition assignment were randomized and counterbalanced. Within each round, clips were presented in a randomized sequence to avoid systematic primacy or recency effects. Condition assignment was balanced across trials using a prioritization algorithm that tracked usage frequencies: for a given clip-arousal target pair, conditions not yet seen at that arousal level were prioritized, followed by those used less frequently overall, and finally those less frequent for that specific target. This ensured that across the experiment, all four conditions appeared equally often per participant and were evenly distributed across arousal levels.

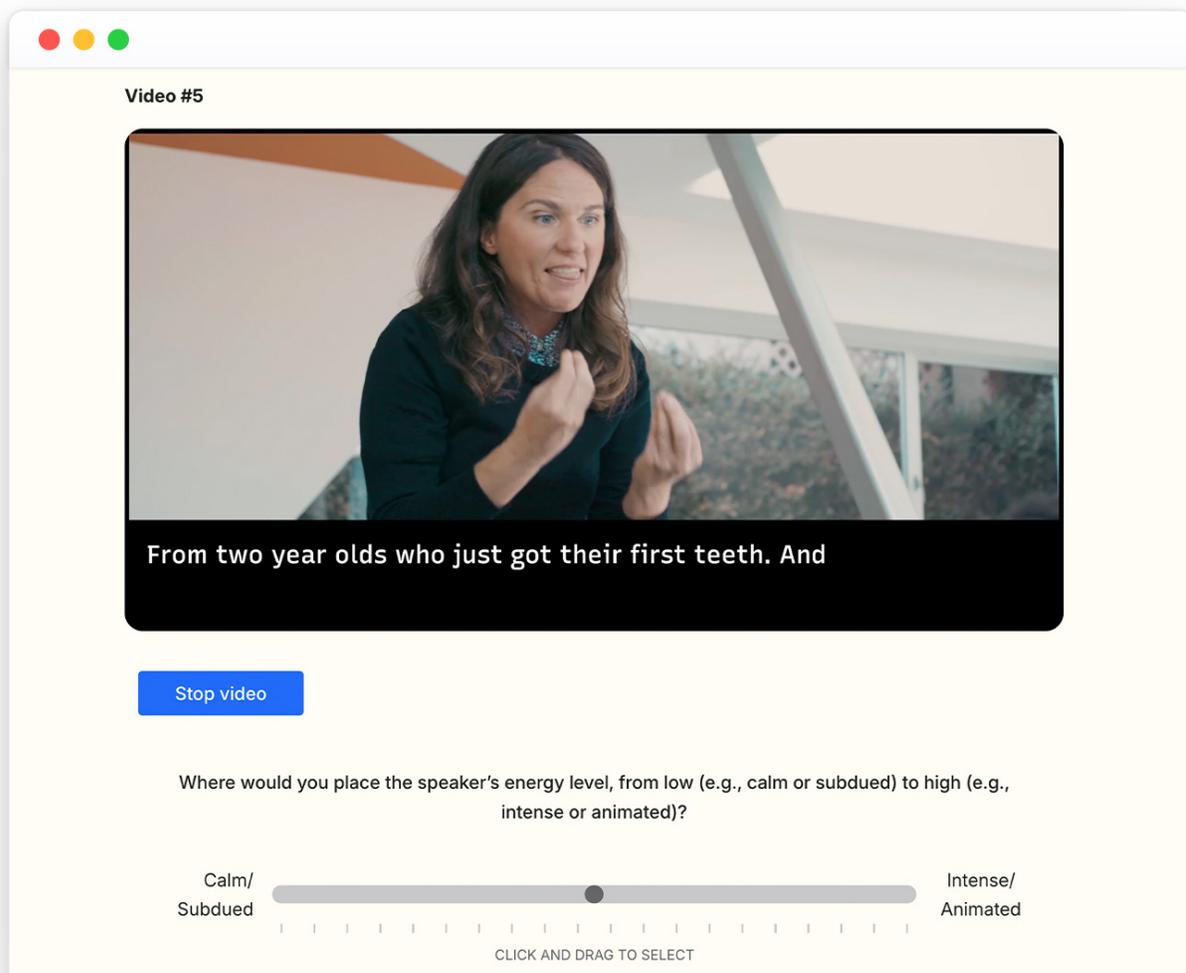


Figure 12.4: Screenshot of the test platform, showing the captioned video (top) and the arousal rating slider (bottom).

12.1.3.1 Outcome

For each trial, participants produced a continuous rating r on the 0–10 arousal scale, representing how intense or animated they perceived the speaker to be. Each clip also had an intended target arousal level T , derived automatically from the speech-emotion recognition model (see Section 12.2). This T value serves as the ground-truth reference for that clip, i.e., the ‘correct’ or expected arousal intensity according to the model. The discrepancy between the perceived and intended level was expressed as an absolute error $E = |r - T|$, with lower E values indicating judgments closer to the intended arousal.

For each participant and condition, we averaged the trial-level errors to obtain a condition-level mean. These values were then entered into a 2×2 repeated-measures analysis (Visuals: on/off; Haptics: on/off; $n = 14$) to test whether the presence of visual or haptic modulations reduced arousal-decoding error.

12.1.4 Experimental Procedure

Sessions were conducted in person by a research assistant proficient in ASL. Procedures were reviewed and approved by our Institutional Review Board. Participants were recruited via local Facebook groups and snowball sampling. Study sessions were held either in a research lab or a library study room. Eligibility required self-identifying as d/Deaf or Hard-of-Hearing. We recruited 14 participants (11 female, 3 male; 9 d/Deaf, 3 Hard-of-Hearing, 2 other: deaf in one ear; deaf & low-vision), mean age 43 ($\sigma=17$). Compensation of \$40 was offered.

At the start, participants performed a brief haptic calibration on the non-dominant wrist: they adjusted the amplifier so the minimum corresponded to a *just-noticeable* vibration and the maximum to a *comfortable* upper bound. Regardless of condition, all videos were presented unmuted. Participants could replay the clip as needed before placing a continuous rating on the slider.

12.2 Findings from Study 6

The target value for each clip represents the intended arousal level derived automatically from the speech signal using Wagner et al. [190]’s speech-emotion recognition model – the same one used throughout this dissertation. (As introduced in Section 12.1.3.1, these target values T are the reference used in computing the participant-model difference $E = |r - T|$.) The model fine-tunes a transformer-based architecture to

predict continuous arousal values from speech. In our pipeline (see Chapter 5), we apply a padding mechanism to obtain per-word values and compute the duration-weighted mean predicted arousal across the clip, treating this mean as the ground-truth reference for analysis.

At each reading, the output of the speech-emotion recognition model was treated as the target arousal value, *i.e.*, the ‘correct’ answer. Participant performance was quantified as the absolute deviation from this target. Because the dependent variable was an ‘error’ score (bounded at zero and potentially skewed), we first assessed whether the distribution met the assumptions of parametric testing. Normality was tested using the Shapiro–Wilk statistic on participant-level mean errors within each condition. Effect sizes are reported as partial η^2 for omnibus tests and as Cohen’s d_z for pairwise contrasts.

We first verified normality (Shapiro–Wilk, all $p > .13$). A repeated-measures ANOVA showed a significant main effect of *Haptics*, $F(1, 13) = 4.82, p = .047, \eta_p^2 = .27$: errors were lower when haptic cues were present ($M = 2.52, SD = 0.63$) than absent ($M = 2.90, SD = 0.99$), a medium-to-large effect with a paired-samples contrast of $t(13) = 2.20, p = .046, d_z = 0.59$. Neither the *Visuals* main effect, $F(1, 13) = 1.58, p = .23, \eta_p^2 = .11$, nor the *Visuals* \times *Haptics* interaction, $F(1, 13) = 1.65, p = .22, \eta_p^2 = .11$, reached significance.

We also conducted exploratory post-hoc pairwise comparisons among the four conditions using paired *t*-tests with Holm correction. Before correction, *Haptics only* tended to yield lower error than *No Visuals/No Haptics*, $t(13) = 2.40, p_{unc} = .032, d_z = 0.64$, and also lower than *Visuals only*, $t(13) = 2.36, p_{unc} = .034, d_z = 0.63$. However, after correcting for multiple comparisons, none of the six contrasts remained significant (all $p_{corr} > .19$). In other words, while the ANOVA revealed a reliable *main effect of Haptics* (collapsing across visuals), no individual pair of conditions could be identified as significantly different once correction was applied.

Condition	Mean Error	SD
No Visuals, No Haptics	2.91	1.01
Visuals only	2.89	1.11
Haptics only	2.35	0.69
Visuals & Haptics	2.69	0.72

Table 12.1: Mean absolute deviation from target arousal (0–10). Lower is better.

In summary, the analyses showed that adding haptic cues reliably reduced arousal-decoding error overall, as reflected in the main effect of *Haptics*. By contrast, we did not observe any significant effect of visual modulations, either alone or in combination with haptics. Exploratory pairwise comparisons suggested

lower errors in the *Haptics only* condition, but these contrasts did not remain significant after correction for multiple tests.

12.3 Discussion of Study 6

12.3.1 Haptics Helped People Decode Arousal Without Extra Visuals

Across conditions, *haptics measurably reduced decoding error* for speakers' arousal levels relative to no haptics, even when presented alone. This shows that the modality is indeed promising as a way of encoding affective information, in particular arousal.

By contrast, we observed *no main effect of visual modulations* (font-weight) on arousal-decoding accuracy. This could reflect our small sample size ($n = 14$) and the subtlety of weight changes, possibly compounded by the fact that our sample skewed older than in our previous studies ($\mu=43, \sigma=17$). Methodological differences from prior evaluations may also have played a role. In Study 3, participants judged static caption images using the EmojiGrid, where they could examine subtle typographic differences at length. Here, in Study 6, captions appeared only briefly and in synchrony with audio (and sometimes haptics). While this dynamic presentation was essential for testing haptics, it may have made fine-grained typographic judgments more difficult.

The contrast with Study 5 is also worth noting. There, font-weight performed better, where, combined with haptics, it improved narrative engagement. Yet those videos were generally longer than in the current study, potentially giving participants more time to 'train' themselves to recognize weight differences, however subtle. Here, by contrast, clips were much shorter, leaving less opportunity for such learning.

A cautious interpretation is that the two modalities may serve complementary purposes: haptics support accurate decoding of affective information, while visuals, given sufficient exposure, can enhance engagement. The current study supports the role of haptics in decoding, but is underpowered to fully assess whether there is a *negative* interaction between the two modalities. Future systems may therefore need to combine them deliberately, while further studies should examine their interaction, including whether *increasing the visual salience* of typographic changes (*e.g.*, larger deltas in weight or shifts in font-size) can improve resiliency to contextual factors (*e.g.*, fast dialogue, brief exposure) and personal factors (*e.g.*, low vision).

12.4 *Limitations*

Because we aimed to obtain single measurements of average arousal from short audio segments, and because these needed a clear ‘target’ mapped onto six arousal bins, it was important that the segments remain brief. This design effectively showed that haptics reduced arousal-perception error, though the video-watching setup was still quite different from more typical experiences. Alternative measurement methods could help further examine how the haptic signal is decoded in more ecologically valid settings.

Our small sample size limits the generalizability of these findings. A larger participant pool could clarify the relative contributions of each condition and reveal whether decoding accuracy varies systematically across the arousal spectrum – for example, whether extreme values (very high or very low) are easier to perceive than mid-range values.

Finally, our ground-truth arousal targets were derived from the continuous predictions of the transformer-based speech-emotion recognition model by Wagner et al. [190], rather than from human annotations. Although the model achieves high agreement with benchmark corpora (concordance correlation coefficients of approximately 0.74 for arousal on MSP-Podcast) and demonstrates robustness across domains and moderate noise conditions, its predictions remain approximations of human perception. The model’s performance can vary by speaker, recording quality, and linguistic domain, and it has been shown to exhibit fairness differences across speakers and genders. Future work could compare model-derived targets with human-labeled data to assess the extent to which such prediction variability affects decoding accuracy.

CHAPTER 13

Conclusion to Part III

In this part, we present and evaluate a novel approach to translate a speaker's arousal levels in the form of haptic signals. These are transmitted to users via a wrist-worn device, providing information about the speaker's emotions that serve as a complement to their transcribed words shown through captions. This method aims to improve the accessibility of spoken communication for individuals who are d/Deaf and Hard-of-Hearing.

Our approach involved designing six distinct patterns by mixing three rhythmic configurations with a low and a high frequency. In Study 4, we tested these patterns with 16 DHH participants. The results showed that the most preferred pattern was a single, short pulse (SSP) per word at a frequency of 75 Hz. Unlike prior work that looked at the design space of visual cues to depict arousal levels, participants' preferences in our study had a higher convergence, suggesting that haptic-feedback can serve as an adequate representation of this emotional dimension in affective captions.

In Study 5, we used the SSP haptic pattern to examine how various combinations of visual cues and haptic feedback influenced the narrative engagement of DHH viewers of audio-visual media. We observed that caption style C_{4V+H} , which integrated both a haptic signal and visual cues to represent arousal, alongside additional visual cues for valence, significantly enhanced engagement compared to a conventional caption style (C_{1B}) devoid of affective information. Additionally, it outperformed another affective captioning style that, based on recommendations from prior work, relied solely on visual cues to convey both arousal and valence (C_{2V}).

Finally, in Study 6 we employed the same SSP haptic pattern, together with font-weight typographic modulations. These were presented either individually or combined, and we measured how well 14 DHH participants were able to decode the inferred arousal values from a speaker in short video clips. We saw

that haptics, by itself, produced a significant reduction in participants' error rate, an effect that neither the baseline, visuals-only, or visuals-and-haptic conditions showed.

Taken together, the studies in Part III show that haptics are a promising addition to affective captions. They not only improve narrative engagement in longer-form viewing (Study 5), but also enhance decoding accuracy of arousal in short clips (Study 6), with consistent user preference for the low-frequency SSP pattern established in Study 4. Visual cues alone were less reliable, but in combination with haptics, they contributed to richer engagement. Overall, Part III demonstrates that affective captions can benefit from a *haptics-first* approach, with haptics ensuring reliable transmission of arousal and their combination with visuals adding engagement, and that future designs should emphasize personalization and salience adjustments to accommodate diverse user needs.

CHAPTER 14

Summary and Contributions

This dissertation looked into the relationship between speech and captions that make it accessible for Deaf and Hard-of-Hearing individuals. Even when transcription is flawless and has little to no latency, we posit that captions might still fall short due to how they do not differentiate vocal expression – words are depicted the same way, regardless of how they were said.

Initially, we investigated this issue by gathering insights from DHH users of teleconferencing software, where communication with hearing partners is typically mediated by captioning systems. Our focus was on understanding their perspectives regarding this gap between the expressive nature of human speech and the flat representation provided by captions.

This exploration prompted us to conduct a subsequent study wherein we experimented with captions that portrayed different combinations of features from speech. In a third study, we explored the design possibilities for these enhanced captions.

A fourth study investigated how a haptic feedback signal echoing dimensions of speech could be used to add to the expressiveness of affective captions. In a follow-up, fifth study, we compared different combinations of visual and haptic cues with a traditional captioning style to measure how each approach affected DHH viewer's narrative engagement with audio-visual content. Lastly, our sixth study investigated how well haptic feedback, by itself or combined with typographic modulations, allowed participants to decode the conveyed arousal levels.

14.1 Contributions

Part I, which included Study 1, looked into the experiences of DHH individuals with automatic captions in remote meetings, identifying gaps in how these systems convey nuances of spoken language. Through eight in-depth interviews, the research highlighted that while caption technology has improved, it fails to capture essential elements like tone and pitch, often leading to misunderstandings and feelings of exclusion among DHH users. The participants' feedback on prototype captioning systems that attempt to include emotional and prosodic cues points to a strong demand for expressive and inclusive captioning solutions. The study shed light on the shortcomings of current technologies, setting the groundwork for my follow-up work by both justifying its need and showing that addressing the absence of paralinguistics in captions is a promising first step.

Part II, which included both Study 2 and 3, looked at both *what* features from speech are most helpful to depict through captions, and *how* to visually do so. In Study 2, DHH participants viewed short videos captioned in four styles – conventional, prosody-only, emotion-only, and prosody & emotion – and, for each, rated emotional clarity, emphasis, legibility, and willingness to use in work/personal settings. Results indicated that captions incorporating emotional cues (either alone or combined with prosody) were more effective in helping participants identify the speaker's emotions and moods compared to conventional captions. However, traditional captions were found to be more legible. Overall, the study showed that embedding emotional information into captions can enhance their communicative value for DHH individuals, though trade-offs in legibility and preferences for use in different settings (work vs. personal meetings) need to be carefully managed.

Beyond demonstrating the feasibility of enhanced captions, the study produced a speech-processing and caption-rendering pipeline that became the technical foundation for the experiments reported in Studies 3–6. It includes both functions for measuring of prosodic cues in US-English, and a machine-learning based method to infer affective cues from speech, along with the visualization of these features through the modulation of typographic parameters in captions.

In Study 3, we explored the preferences of DHH participants for caption styles that depict valence and arousal. We were able to identify nine typographic parameters that can be applied to captions as a way of conveying bipolar scale such as valence and arousal. We measured DHH participants' preferences for the use of these nine styles for depicting either valence or arousal, at first, and then both valence and arousal. The top-performing choices of the latter were then evaluated for their EASE OF READING,

LOW DISTRACTION, INTUITIVENESS and CLARITY OF EMOTIONAL REPRESENTATION. We offer two designs as recommendations for affective captioning applications, namely, using font-color for valence and font-weight for arousal, or using font-color for valence and font-size for arousal. The former is to be preferred in applications where legibility is at a premium.

Methodologically, Study 3 offered an exploration of the use of Best-worst scaling for capturing explicit and implicit preferences, paired with an ELO-rating system for analyzing data and ranking preference choices. We also showcased the use of EmojiGrid measures for affective captioning studies.

Part III includes Study 4, where we explored how and whether haptic feedback can complement the visuals-only approaches we had previously developed towards more inclusive and engaging captioning systems. To do so, we introduced a per-word, amplitude-modulated vibration stream that mirrors a speaker's arousal levels and can run alongside visuals-only affective captions. The implementation includes a novel word-timed haptic synthesis pipeline, along with methods to incorporate it within an inexpensive wrist-worn actuator setup suitable for lab studies.

Using best-worst scaling with TrueSkill ranking, we identified a preferred option among six combinations of rhythmic and frequency patterns: a *single short pulse* per word at 75 Hz, with amplitude carrying arousal. We also documented comfort and distraction trade-offs (*e.g.*, user discomfort at 250 Hz) and established a simple calibration procedure to balance device response and perceived intensity.

In Study 5, we applied this top-performing choice, together with the winning set of visual properties coming out on top from Study 3, and examined how their varying combinations influenced viewers' engagement with audiovisual content. The condition combining both visual and haptic modulations significantly outperformed a baseline and visuals-only condition in terms of Narrative Engagement.

Finally, in Study 6, we tested how well our proposed haptic feedback pattern, whether combined with visuals or not, allowed participants to decode the varying arousal levels inferred from speech. Haptics had a significant effect in helping participants more accurately interpret these values, whereas visuals alone or in combination with haptics did not yield a comparable effect.

As such, combined findings from Study 3 through 6 led to four actionable design guidance:

1. Depict valence with font-color and arousal with font-weight when legibility is a priority;
2. Pair visuals with a single-pulse 75 Hz haptic track for arousal when optimizing for engagement;

3. Avoid dense, multi-pulse rhythms, especially when coupled with high-frequency (250 Hz) vibrations, due to both distraction and comfort issues;
4. Consider thresholding and personalization (*e.g.*, vibrating only salient arousal peaks) to reduce fatigue in fast speech or long-form content.

Lastly, our work provided methodological advances for caption research. Namely, we were able to adapt the Narrative Engagement instrument to captioned video comparisons, while also showing how BWS → TrueSkill can efficiently down-select haptic patterns, and provided a reproducible, word-synchronous toolchain that aligns ASR, affect inference, visual modulation, and haptic output.

14.2 Publications

The completed research presented in this dissertation has led to three publications at the ACM CHI Conference on Human Factors in Computing Systems, the premier international conference on Human-Computer Interaction:

CHI 2023

The research presented at CHI'23, detailed in Chapters 2 and 5, comprised two parts. The first was an interview-based study that investigated the challenges faced by DHH users of captioned teleconferencing applications and identified potential improvements. The second part evaluated various captioning prototypes, assessing the effectiveness of different combinations of features from speech for enhancing DHH participants' comprehension of speakers [53].

CHI 2024

The research presented at CHI'24, detailed in Chapters 6, explored the design space of affective captions through a three-phase study. The first phase measured preferences among DHH users for typographic features like color, boldness, and size. The second phase tested combinations of these preferred styles for effectively portraying both valence and arousal. The final phase compared these combinations against each other and a non-styled baseline, focusing on criteria such as readability and emotional clarity. The study culminates with a set of design recommendations for affective captions [55]. *This paper received an Honorable Mention (top 5%).*

CHI 2025

The research presented at CHI'25, detailed in Chapter 10, introduced a multimodal approach to affective captions that encodes arousal as per-word, amplitude-modulated vibrations that, as we did, can be delivered via a wrist-worn device alongside visual cues. In a formative study, we compared six rhythm \times frequency patterns and identified a preferred option for DHH users: a single short pulse per word at 75 Hz. A follow-up within-subjects study showed that combining this haptic channel with visuals significantly increased narrative engagement compared to both a conventional baseline and a visuals-only affective caption style, supporting a design recommendation to pair visual valence with haptic arousal [54].

Related, not included as a chapter. Our ASSETS 2025 paper – *CuCap: Comparative Analysis of Customized Captioning between North American and South Korean d/Deaf and Hard-of-Hearing Users* [50] – extends the findings from Study 3 by evaluating a customizable affective/prosodic captioning interface (*CuCap*) with 49 DHH participants in North America and South Korea. Emotion visualizations were consistently preferred across both groups, whereas preferences for prosodic mappings diverged: Korean participants selected features often excluded by North American participants. These cross-cultural differences motivate flexible, user-configurable caption systems.

14.2.1 *Additional Publications*

Beyond the work included as dissertation chapters, I have also contributed to several other projects during my Ph.D., which explore related themes of accessibility, video comprehension, and interactive learning systems:

ASSETS 2022

Support in the Moment: Benefits and Use of Video-Span Selection and Search for Sign-Language Video Comprehension among ASL Learners [82]. Introduced a prototype that lets learners select spans of signing videos to retrieve dictionary results in situ, improving translation quality and reducing workload compared to conventional tools. *This paper received a Best Paper nomination (top 6%).*

TACCESS

Exploring the Benefits and Applications of Video-Span Selection and Search for Real-Time Support

in Sign Language Video Comprehension among ASL Learners [83]. Extended the ASSETS 2022 work with three studies, showing span-selection improved comprehension speed and accuracy, and offering design recommendations for sign-language learning systems.

CHI 2024

Designing and Evaluating an Advanced Dance Video Comprehension Tool with In-Situ Move Identification Capabilities [84]. Extended the Wizard-of-Oz prototype used for ASL learning for the context of dance education, showing automatic move identification enhanced note-taking and reduced workload, pointing to opportunities for AI-supported movement learning. *This paper received an Honorable Mention (top 5%)*.

ASSETS 2025

CapTune: Adapting Non-Speech Captions with Anchored Generative Models [89]. Presented a customizable caption system balancing creator intent and viewer control. Evaluations with creators and DHH viewers showed enhanced engagement and revealed trade-offs between expressiveness and cognitive load.

14.3 Future Directions

The visual and haptic approaches discussed here were able to produce measurable improvements in factors such as recognition of emotions in speech and engagement with audiovisual content. Work still needs to be done to better understand whether affective captions improve *understanding*. While enabling users to perceive affective information through visuals and haptics is an important first step, their broader impacts on how viewers make sense of spoken content remain an open question. Future work could examine both the effects of affective captions in comprehension and memory metrics, for when the material is prerecorded, or conversational outcomes, when captions are used in live meetings – for instance, whether users respond differently to questions when presented through affective captions.

Studying users' reactions to longer videos, or even videos presented over time in extended longitudinal studies, could bring important insights into how the proposed visual and haptic modulations behave as users get used to them. It is plausible that, through learning, the decoding and immersion effects we observed could improve over time. At the same time, distraction and fatigue remain plausible trade-offs,

which may call for more discrete visual/haptic modulations or adjustments to the algorithmic feature-selection process (*e.g.*, setting thresholds below which cues are ignored).

In parallel, these studies could incorporate videos that stress the system with more challenging speaking scenarios, such as those involving multiple speakers, or videos where other non-speech information cues like speaker identification labels, music and sound effects descriptions, lyrics, etc, compete with the affective cues. Viewing settings could also be expanded, including different form factors and contexts, *e.g.*, mobile phones, larger displays, watching in public, with other viewers (DHH or not), etc.

A key takeaway throughout many of our studies was that user-personalization is a key consideration when designing captioning systems. Imagining a complex richer feature set raises complex interface and experience challenges, given how these systems will need to convey additional dimensions of non-speech information. How can users make sense of these complex, multi-dimensional settings in a way that is intuitive but does not force them to forfeit agency?

14.4 *Concluding Thoughts*

This dissertation advances the understanding of communication barriers faced by the DHH community that, while subtle, can be significant. By exploring the representation of nuanced dimensions of speech through captions, using both visual and haptic mediums, we aim to improve how DHH individuals can communicate with hearing peers and, more generally, engage with audio-visual content. We employ a novel combination of methods and approaches that have allowed us to explore hard-to-measure dimensions of perception, and we intend these to inspire other researchers working in similar sets of problems. Finally, the approaches discussed here are not exhaustive, and we expect they will serve as a foundation for further exploratory and applied research, particularly in developing more inclusive technologies.

Bibliography

- [1] Joshua M. Ackerman, Christopher C. Nocera, and John A. Bargh. 2010. Incidental Haptic Sensations Influence Social Judgments and Decisions. *Science* 328, 5986 (June 2010), 1712–1715. <https://doi.org/10.1126/science.1189993>
- [2] Elvar Atli Ævarsson, Thórhildur Ásgeirsdóttir, Finnur Pind, Árni Kristjánsson, and Runar Unnthorsson. 2022. Vibrotactile Threshold Measurements at the Wrist Using Parallel Vibration Actuators. *ACM Trans. Appl. Percept.* 19, 3, Article 10 (sep 2022), 11 pages. <https://doi.org/10.1145/3529259>
- [3] Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How Might We Create Better Benchmarks for Speech Recognition?. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*. Association for Computational Linguistics, Online, 22–34. <https://doi.org/10.18653/v1/2021.bppf-1.4>
- [4] Akshita, Harini Alagarai Sampath, Bipin Indurkhya, Eunhwa Lee, and Yudong Bae. 2015. Towards Multimodal Affective Feedback: Interaction between Visual and Haptic Modalities. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) (*CHI '15*). Association for Computing Machinery, New York, NY, USA, 2043–2052. <https://doi.org/10.1145/2702123.2702288>
- [5] Ali Alaraj, Fady T. Charbel, Daniel Birk, Mathew Tobin, Cristian Luciano, Pat P. Banerjee, Silvio Rizzi, Jeff Sorenson, Kevin Foley, Konstantin Slavin, and Ben Roitberg. 2013. Role of Cranial and Spinal Virtual and Augmented Reality Simulation Using Immersive Touch Modules in Neurosurgical Training. *Neurosurgery* 72, Supplement 1 (Jan. 2013), A115–A123. <https://doi.org/10.1227/neu.0b013e3182753093>

- [6] 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 Communications Association, Poznan, Poland, 13–16.
- [7] Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2021. Effect of Occlusion on Deaf and Hard of Hearing Users' Perception of Captioned Video Quality. In *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments*, Margherita Antona and Constantine Stephanidis (Eds.). Springer International Publishing, Cham, 202–220.
- [8] 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 *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>
- [9] 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>
- [10] 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.
- [11] Rasmus Andersson. 2023. The Inter typeface family. <https://rsms.me/inter/>. [Online; accessed 22-November-2023].
- [12] Carl Armon, Dan Glisson, and Larry Goldberg. 1992. How Closed Captioning in the U.S. Today can Become the Advanced Television Captioning System of Tomorrow. *SMPTE Journal* 101, 7 (1992), 495–498. <https://doi.org/10.5594/J02244>
- [13] Salvatore Attardo, Manuela Maria Wagner, and Eduardo Urios-Aparisi. 2011. Prosody and humor. *Pragmatics & Cognition* 19, 2 (2011), 189–201.

- [14] Plínio A. Barbosa. 2019. *Prosódia*. Parábola Editorial.
- [15] 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>
- [16] Sofie Beier, Sam Berlow, Esat Boucaud, Zoya Bylinskii, Tianyuan Cai, Jenae Cohn, Kathy Crowley, Stephanie L. Day, Tilman Dingler, Jonathan Dobres, Jennifer Healey, Rajiv Jain, Marjorie Jordan, Bernard Kerr, Qisheng Li, Dave B. Miller, Susanne Nobles, Alexandra Papoutsaki, Jing Qian, Tina Rezvanian, Shelley Rodrigo, Ben D. Sawyer, Shannon M. Sheppard, Bram Stein, Rick Treitman, Jen Vanek, Shaun Wallace, and Benjamin Wolfe. 2021. Readability Research: An Interdisciplinary Approach. *CoRR* abs/2107.09615 (2021), 85 pages. arXiv:2107.09615 <https://arxiv.org/abs/2107.09615>
- [17] S. J. Bensmaïa, Y. Y. Leung, S. S. Hsiao, and K. O. Johnson. 2005. Vibratory Adaptation of Cutaneous Mechanoreceptive Afferents. *Journal of Neurophysiology* 94, 5 (Nov. 2005), 3023–3036. <https://doi.org/10.1152/jn.00002.2005>
- [18] 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>
- [19] Larwan Berke, Khaled Albusays, Matthew Seita, and Matt Huenerfauth. 2019. Preferred Appearance of Captions Generated by Automatic Speech Recognition for Deaf and Hard-of-Hearing Viewers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (*CHI EA '19*). Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3290607.3312921>
- [20] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. 2017. Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One Meetings. 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, 155–164. <https://doi.org/10.1145/3132525.3132541>
- [21] 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>
- [22] 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.
- [23] Natesh M Bhat. 2021. Text-to-speech x-platform¶. <https://pyttsx3.readthedocs.io/en/latest/>
- [24] Alexander Bick, Adam Blandin, and Karel Mertens. 2023. Work from Home before and after the COVID-19 Outbreak. *American Economic Journal: Macroeconomics* 15, 4 (Oct. 2023), 1–39. <https://doi.org/10.1257/mac.20210061>
- [25] Samantha W. Bindman, Lisa M. Castaneda, Mike Scanlon, and Anna Cechony. 2018. Am I a Bunny? The Impact of High and Low Immersion Platforms and Viewers' Perceptions of Role on Presence, Narrative Engagement, and Empathy during an Animated 360° Video. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI '18*). Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3173574.3174031>
- [26] Frank Biocca, Jin Kim, and Yung Choi. 2001. Visual touch in virtual environments: An exploratory study of presence, multimodal interfaces, and cross-modal sensory illusions. *Presence: Teleoperators & Virtual Environments* 10, 3 (2001), 247–265.
- [27] Jeffrey R. Blum, Jessica R. Cauchard, and Jeremy R. Cooperstock. 2020. Habituation to Pseudo-Ambient Vibrotactile Patterns for Remote Awareness. In *2020 IEEE Haptics Symposium (HAPTICS)*. IEEE, Washington, D.C., USA, 657–663. <https://doi.org/10.1109/haptics45997.2020.ras.hap20.153.550dbcba>
- [28] 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>
- [29] 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>

- [30] Paul Boersma. 2006. Praat: doing phonetics by computer. <http://www.praat.org/>. Accessed on August 24, 2022.
- [31] Jason T. Bowey and Regan L. Mandryk. 2017. Those are not the Stories you are Looking For: Using Text Prototypes to Evaluate Game Narratives Early. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Amsterdam, The Netherlands) (*CHI PLAY '17*). Association for Computing Machinery, New York, NY, USA, 265–276. <https://doi.org/10.1145/3116595.3116636>
- [32] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [33] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
- [34] Rick Busselle and Helena Bilandzic. 2008. Fictionality and Perceived Realism in Experiencing Stories: A Model of Narrative Comprehension and Engagement. *Communication Theory* 18, 2 (May 2008), 255–280. <https://doi.org/10.1111/j.1468-2885.2008.00322.x>
- [35] Rick Busselle and Helena Bilandzic. 2009. Measuring Narrative Engagement. *Media Psychology* 12, 4 (Nov. 2009), 321–347. <https://doi.org/10.1080/15213260903287259>
- [36] 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>
- [37] Tianyuan Cai, Shaun Wallace, Tina Rezvani, 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>
- [38] João Couceiro e Castro, Pedro Martins, Ana Boavida, and Penousal Machado. 2019. Máquina de Ouvir-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*. Association for Computing Machinery, Braga, Portugal, 1–8.

- [39] Anna C. Cavender, Jeffrey P. Bigham, and Richard E. Ladner. 2009. ClassInFocus: Enabling Improved Visual Attention Strategies for Deaf and Hard of Hearing Students. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, Pennsylvania, USA) (*Assets '09*). Association for Computing Machinery, New York, NY, USA, 67–74. <https://doi.org/10.1145/1639642.1639656>
- [40] 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>
- [41] 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.
- [42] 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.
- [43] Jonathan Cohen. 2018. *Defining Identification: A Theoretical Look at the Identification of Audiences With Media Characters*. Routledge, 253–272. <https://doi.org/10.4324/9781315164441-14>
- [44] Quentin Consigny, Nathan Ouvrai, Arthur Paté, Claudia Fritz, and Jean-Loïc Le Carrou. 2023. Vibrotactile Thresholds on the Wrist for Vibrations Normal to the Skin. *IEEE Transactions on Haptics* (2023), 1–6. <https://doi.org/10.1109/TOH.2023.3275185>
- [45] Michael Cooper. 2021. W3C Accessibility Guidelines (WCAG) 3.0. <https://www.w3.org/TR/wcag-3.0/>
- [46] 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>
- [47] Wellington da Silva, Plinio Almeida Barbosa, and Åsa Abelin. 2016. Cross-Cultural and Cross-Linguistic Perception of Authentic Emotions through Speech: An Acoustic-Phonetic Study with Brazilian and Swedish Listeners. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada* 32, 2 (Aug. 2016), 449–480. <https://doi.org/10.1590/0102-445003263701432483>

- [48] 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.
- [49] 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>
- [50] Caluã de Lacerda Pataca, SooYeon Ahn, Suhyeon Yoo, JooYeong Kim, Khai N. Truong, Jin-Hyuk Hong, Roshan L. Peiris, and Matt Huenerfauth. 2025. CuCap: Comparative Analysis of Customized Captioning between North American and South Korean d/Deaf and Hard-of-Hearing Users. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25)* (Denver, CO, USA). Association for Computing Machinery, New York, NY, USA, 1–21. <https://doi.org/10.1145/3663547.3746400>
- [51] 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*. Association for Computing Machinery, Cagliari, Italy, 139–143.
- [52] 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>
- [53] 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. 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 831, 15 pages. <https://doi.org/10.1145/3544548.3581511>
- [54] Caluã de Lacerda Pataca, Saad Hassan, Lloyd May, Michelle M Olson, Toni D'aurio, Roshan L Peiris, and Matt Huenerfauth. 2025. Tactile Emotions: Multimodal Affective Captioning with Haptics Improves Narrative Engagement for d/Deaf and Hard-of-Hearing Viewers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 68, 17 pages. <https://doi.org/10.1145/3706598.3713304>

- [55] Caluã de Lacerda Pataca, Saad Hassan, Nathan Tinker, Roshan Lalintha 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* (, Honolulu, HI, USA,) (*CHI '24*). Association for Computing Machinery, New York, NY, USA, Article 899, 17 pages. <https://doi.org/10.1145/3613904.3642258>
- [56] 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>
- [57] João Antônio de Moraes and Albert Rilliard. 2016. Prosody and Emotion in Brazilian Portuguese. In *Issues in Hispanic and Lusophone Linguistics*, Meghan E. Armstrong, Nicholas Henriksen, and Maria del Mar Vanrell (Eds.). Vol. 6. John Benjamins Publishing Company, Amsterdam, 135–152. <https://doi.org/10.1075/ihl1.6.07mor>
- [58] Jorge dos Reis. 2014. Speechant: Chanting & Speeching: Sistema de notação tipográfica para a educação de adultos. *Matéria Prima* 2, 3 (2014).
- [59] J. dos Reis and V. Hazan. 2011. Speechant: a vowel notation system to teach English pronunciation. *ELT Journal* 66, 2 (June 2011), 156–165. <https://doi.org/10.1093/elt/ccr019>
- [60] Lisa Elliot, Michael Stinson, James Mallory, Donna Easton, and Matt Huenerfauth. 2016. Deaf and Hard of Hearing Individuals' Perceptions of Communication with Hearing Colleagues in Small Groups. 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, 271–272. <https://doi.org/10.1145/2982142.2982198>
- [61] Arpad E Elo. 1978. *The rating of chessplayers, past and present*. Arco Pub., New York.
- [62] Lisa A. Feldman. 1995. Variations in the Circumplex Structure of Mood. *Personality and Social Psychology Bulletin* 21, 8 (Aug. 1995), 806–817. <https://doi.org/10.1177/0146167295218003>
- [63] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying Bias in Automatic Speech Recognition. <https://doi.org/10.48550/ARXIV.2103.15122>
- [64] Leah Findlater, Bonnie Chinh, Dhruv Jain, Jon Froehlich, Raja Kushalnagar, and Angela Carey Lin. 2019. Deaf and Hard-of-hearing Individuals' Preferences for Wearable and Mobile Sound Awareness Technologies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Sys-*

- tems* (Glasgow, Scotland Uk) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300276>
- [65] Harvey Fletcher and Wilden A Munson. 1933. Loudness, its definition, measurement and calculation. *Bell System Technical Journal* 12, 4 (1933), 377–430.
- [66] Mark D. Fletcher. 2021. Can Haptic Stimulation Enhance Music Perception in Hearing-Impaired Listeners? *Frontiers in Neuroscience* 15 (2021). <https://doi.org/10.3389/fnins.2021.723877>
- [67] Mark D Fletcher, Amatullah Hadeedi, Tobias Goehring, and Sean R Mills. 2019. Electro-haptic enhancement of speech-in-noise performance in cochlear implant users. *Scientific Reports* 9, 1 (2019), 11428.
- [68] Alejandro Flores Ramones and Marta Sylvia del Rio-Guerra. 2023. Recent Developments in Haptic Devices Designed for Hearing-Impaired People: A Literature Review. *Sensors* 23, 6 (2023). <https://doi.org/10.3390/s23062968>
- [69] 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>
- [70] Synnaeve Gabriel. 2022. WER are we? An attempt at tracking states of the art(s) and recent results on speech recognition. https://github.com/syhw/wer_are_we
- [71] E. Gatti, G. Caruso, M. Bordegoni, and C. Spence. 2013. Can the feel of the haptic interaction modify a user's emotional state?. In *2013 World Haptics Conference (WHC)*. IEEE, 247–252. <https://doi.org/10.1109/whc.2013.6548416>
- [72] William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a Resource for Design. 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, 233–240. <https://doi.org/10.1145/642611.642653>
- [73] Sandrine Gil and Ludovic Le Bigot. 2016. Colour and emotion: children also associate red with negative valence. *Developmental science* 19, 6 (2016), 1087–1094.
- [74] Daniel T Gilbert. 1991. How mental systems believe. *American psychologist* 46, 2 (1991), 107.

- [75] Steven Goodman, Susanne Kirchner, Rose Guttman, Dhruv Jain, Jon Froehlich, and Leah Findlater. 2020. Evaluating Smartwatch-based Sound Feedback for Deaf and Hard-of-hearing Users Across Contexts. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '20*). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376406>
- [76] 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>
- [77] Ana Guerberof-Arenas, Joss Moorkens, and David Orrego-Carmona. 2024. “A Spanish version of EastEnders”: a reception study of a telenovela subtitled using MT. *The Journal of Specialised Translation* 141 (Jan. 2024), 230–254. <https://doi.org/10.26034/cm.jostrans.2024.4724>
- [78] Ana Guerberof-Arenas and Antonio Toral. 2024. To be or not to be: A translation reception study of a literary text translated into Dutch and Catalan using machine translation. *Target* (April 2024), 215–244. <https://doi.org/10.1075/target.22134.gue>
- [79] 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>
- [80] Awni Hannun. 2021. The History of Speech Recognition to the Year 2030. arXiv:2108.00084 [cs.CL]
- [81] 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](https://doi.org/10.1016/s0166-4115(08)62386-9)
- [82] Saad Hassan, Akhter Al Amin, Caluã de Lacerda Pataca, Diego Navarro, Alexis Gordon, Sooyeon Lee, and Matt Huenerfauth. 2022. Support in the Moment: Benefits and use of video-span selection and search for sign-language video comprehension among ASL learners. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) (*ASSETS '22*). Association for Computing Machinery, New York, NY, USA, Article 29, 14 pages. <https://doi.org/10.1145/3517428.3544883>

- [83] Saad Hassan, Caluã de Lacerda Pataca, Akhter Al Amin, Laleh Nourian, Diego Navarro, Sooyeon Lee, Alexis Gordon, Matthew Watkins, Garreth W. Tigwell, and Matt Huenerfauth. 2024. Exploring the Benefits and Applications of Video-Span Selection and Search for Real-Time Support in Sign Language Video Comprehension among ASL Learners. *ACM Trans. Access. Comput.* 17, 3, Article 14 (Oct. 2024), 35 pages. <https://doi.org/10.1145/3690647>
- [84] Saad Hassan, Caluã De Lacerda Pataca, Laleh Nourian, Garreth W. Tigwell, Briana Davis, and Will Zhenya Silver Wagman. 2024. Designing and Evaluating an Advanced Dance Video Comprehension Tool with In-situ Move Identification Capabilities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '24*). Association for Computing Machinery, New York, NY, USA, Article 888, 19 pages. <https://doi.org/10.1145/3613904.3642710>
- [85] 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>
- [86] 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/>
- [87] Kristina Höök, Anna Ståhl, Petra Sundström, and Jarmo Laaksolahti. 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>
- [88] Jiaxiong Hu, Qian Yao Xu, Limin Paul Fu, and Yingqing Xu. 2019. Emojilization: An Automated Method For Speech to Emoji-Labeled Text. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (*CHI EA '19*). Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3290607.3313071>
- [89] Jeremy Zhengqi Huang, Caluã de Lacerda Pataca, Liang-Yuan Wu, and Dhruv Jain. 2025. CapTune: Adapting Non-Speech Captions with Anchored Generative Models. In *Proceedings of the 27th Inter-*

- national ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25)*. Association for Computing Machinery, New York, NY, USA, 1–18. <https://doi.org/10.1145/3663547.3746346>
- [90] Gary Hustwit. 2015. *Helvetica/Objectified/Urbanized: The Complete Interviews: The Design Trilogy Interviews*. Versions Publishing, London, UK.
- [91] Dhruv Jain, Brendon Chiu, Steven Goodman, Chris Schmandt, Leah Findlater, and Jon E. Froehlich. 2020. Field Study of a Tactile Sound Awareness Device for Deaf Users. In *Proceedings of the 2020 ACM International Symposium on Wearable Computers (Virtual Event, Mexico) (ISWC '20)*. Association for Computing Machinery, New York, NY, USA, 55–57. <https://doi.org/10.1145/3410531.3414291>
- [92] Martina Jakesch and Claus-Christian Carbon. 2012. The Mere Exposure Effect in the Domain of Haptics. *PLoS ONE* 7, 2 (Feb. 2012), e31215. <https://doi.org/10.1371/journal.pone.0031215>
- [93] Jinkyu Jang, Jinwook Kim, Hyeonsik Shin, Hajung Aum, and Jinwoo Kim. 2016. Effects of Temporal Format of Everyday Video on Narrative Engagement and Social Interactivity. *Interacting with Computers* 28, 6 (Jan. 2016), 718–736. <https://doi.org/10.1093/iwc/iwv043>
- [94] Carl J Jensema, Ramalinga Sarma Danturthi, and Robert Burch. 2000. Time spent viewing captions on television programs. *American annals of the deaf* 145, 5 (2000), 464–468.
- [95] Domicile 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.
- [96] 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>
- [97] Saba Kawas, George Karalis, Tzu Wen, and Richard E. Ladner. 2016. Improving Real-Time Captioning Experiences for Deaf and Hard of Hearing Students. 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, 15–23. <https://doi.org/10.1145/2982142.2982164>

- [98] Hyunju Kim, Yan Tao, Chuanrui Liu, Yuzhuo Zhang, and Yuxin Li. 2023. Comparing the Impact of Professional and Automatic Closed Captions on Video-Watching Experience. 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 74, 6 pages. <https://doi.org/10.1145/3544549.3585634>
- [99] 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>
- [100] 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>
- [101] 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>
- [102] Maria Kraxenberger, Winfried Menninghaus, Anna Roth, and Mathias Scharinger. 2018. Prosody-Based Sound-Emotion Associations in Poetry. *Frontiers in Psychology* 9 (2018). <https://doi.org/10.3389/fpsyg.2018.01284>
- [103] Jan-Louis Kruger, María T. Soto-Sanfiel, Stephen Doherty, and Ronny Ibrahim. 2016. Towards a cognitive audiovisual translatology. In *Reembedding Translation Process Research*. John Benjamins Publishing Company, Amsterdam, 171–194. <https://doi.org/10.1075/btl.128.09kru>
- [104] Christof Kuhbandner and Reinhard Pekrun. 2013. Joint effects of emotion and color on memory. *Emotion* 13, 3 (2013), 375.
- [105] Raja S. Kushalnagar, Gary W. Behm, Joseph S. Stanislow, and Vasu Gupta. 2014. Enhancing caption accessibility through simultaneous multimodal information: visual-tactile captions. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility* (Rochester, New York, USA) (*ASSETS '14*). Association for Computing Machinery, New York, NY, USA, 185–192. <https://doi.org/10.1145/2661334.2661381>

- [106] Raja S. Kushalnagar and Christian Vogler. 2020. Teleconference Accessibility and Guidelines for Deaf and Hard of Hearing Users. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) (ASSETS '20). Association for Computing Machinery, New York, NY, USA, Article 9, 6 pages. <https://doi.org/10.1145/3373625.3417299>
- [107] Marc Wilhelm Küster. 2016. Writing Beyond the Letter. *Tijdschrift voor Mediageschiedenis* 19, 2 (12 2016).
- [108] Jari Laarni, Niklas Ravaja, Timo Saari, Saskia Böcking, Tilo Hartmann, and Holger Schramm. 2015. Ways to Measure Spatial Presence: Review and Future Directions. In *Immersed in Media*. Springer International Publishing, 139–185. https://doi.org/10.1007/978-3-319-10190-3_8
- [109] Walter S. Lasecki, Raja Kushalnagar, and Jeffrey P. Bigham. 2014. Helping Students Keep up with Real-Time Captions by Pausing and Highlighting. In *Proceedings of the nth Web for All Conference* (Seoul, Korea) (*W4A '14*). Association for Computing Machinery, New York, NY, USA, Article 39, 8 pages. <https://doi.org/10.1145/2596695.2596701>
- [110] Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. 2023. From User Perceptions to Technical Improvement: Enabling People Who Stutter to Better Use Speech Recognition. 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 361, 16 pages. <https://doi.org/10.1145/3544548.3581224>
- [111] Colin Lea, Dianna Yee, Jaya Narain, Zifang Huang, Lauren Tooley, Jeffrey P. Bigham, and Leah Findlater. 2023. Latent Phrase Matching for Dysarthric Speech. arXiv:2306.05446 [eess.AS]
- [112] Daniel G Lee, Deborah I Fels, and John Patrick Udo. 2007. Emotive captioning. *Computers in Entertainment (CIE)* 5, 2 (2007), 11.
- [113] Heungsub Lee. 2018. Computing Your Skill. <https://trueskill.org/> [Online; accessed 29-April-2023].
- [114] 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.

- [115] Jason Livingston. 2012. Closed Captioning Challenges for IP Video Delivery. In *The 2012 Annual Technical Conference & Exhibition*. SMPTE, Hollywood, CA, USA, 1–9.
- [116] 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 *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (Virtual Event, Greece) (ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 12, 6 pages. <https://doi.org/10.1145/3373625.3417300>
- [117] Jérôme Louradour. 2023. whisper-timestamped. <https://github.com/linto-ai/whisper-timestamped>.
- [118] Catarina Maçãs, David Palma, and Artur Rebelo. 2019. TypEm: A Generative Typeface That Represents the Emotion of the Text. 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 5, 10 pages. <https://doi.org/10.1145/3359852.3359874>
- [119] Karon E MacLean. 2008. Haptic interaction design for everyday interfaces. *Reviews of Human Factors and Ergonomics* 4, 1 (2008), 149–194.
- [120] Fiona Macpherson. 2018. *Sensory Substitution and Augmentation: An Introduction*. British Academy, 1–42. <https://doi.org/10.5871/bacad/9780197266441.003.0001>
- [121] Tim Mahrt. 2022. PraatIO. <https://github.com/timmahrt/praatIO>. Accessed on August 3, 2022.
- [122] Sabrina Malik, Jonathan Aitken, and Judith Kelly Waalen. 2009. Communicating emotion with animated text. *visual communication* 8, 4 (2009), 469–479.
- [123] 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>
- [124] Lloyd May, Rabia Malik, and AnnMarie Thomas. 2024. Co-Designing Haptic Instruments With Deaf and Hard-of-Hearing Children. (2024). <https://doi.org/10.5281/ZENODO.13904780>

- [125] Lloyd May, Sarah Miller, Sehuam Bakri, Lorna C Quandt, and Melissa Malzkuhn. 2023. Designing Access in Sound Art Exhibitions: Centering Deaf Experiences in Musical Thinking. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–8.
- [126] Lloyd May, So Yeon Park, and Jonathan Berger. 2023. Enhancing Non-Speech Information Communicated in Closed Captioning Through Critical Design. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (<conf-loc>, <city>New York</city>, <state>NY</state>, <country>USA</country>, </conf-loc>) (*ASSETS '23*). Association for Computing Machinery, New York, NY, USA, Article 16, 14 pages. <https://doi.org/10.1145/3597638.3608398>
- [127] Gretchen McCulloch. 2019. *Because Internet: Understanding the New Rules of Language* (1st ed.). Riverhead Books, New York. Kindle version.
- [128] 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>
- [129] Emma J McDonnell, Soo Hyun Moon, Lucy Jiang, Steven M. Goodman, Raja Kushalnagar, Jon E. Froehlich, and Leah Findlater. 2023. “Easier or Harder, Depending on Who the Hearing Person Is”: Codesigning Videoconferencing Tools for Small Groups with Mixed Hearing Status. 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 780, 15 pages. <https://doi.org/10.1145/3544548.3580809>
- [130] Aleksandro R Meireles, João Paulo Tozetti, and Rogério R Borges. 2010. Speech rate and rhythmic variation in Brazilian Portuguese. In *Speech Prosody 2010-Fifth International Conference*. International Speech Communication Association (ISCA), Chicago, USA, 1–4.
- [131] Sebastian Merchel and M Ercan Altinsoy. 2018. Auditory-tactile experience of music. *Musical Haptics* (2018), 123–148.
- [132] Chris Mikul. 2014. *Caption quality: Approaches to standards and measurement*. Media Access Australia, Sydney, Australia.

- [133] Kouta Minamizawa, Yasuaki Kakehi, Masashi Nakatani, Soichiro Mihara, and Susumu Tachi. 2012. TECHTILE toolkit: a prototyping tool for design and education of haptic media. In *Proceedings of the 2012 Virtual Reality International Conference (Laval, France) (VRIC '12)*. Association for Computing Machinery, New York, NY, USA, Article 26, 2 pages. <https://doi.org/10.1145/2331714.2331745>
- [134] Anant Mittal, Meghna Gupta, Roshni Poddar, Tarini Naik, Seethalakshmi Kuppuraj, James Fogarty, Pratyush Kumar, and Mohit Jain. 2023. Jod: Examining Design and Implementation of a Videoconferencing Platform for Mixed Hearing Groups. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (<conf-loc>, <city>New York</city>, <state>NY</state>, <country>USA</country>, </conf-loc>) (ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 43, 18 pages. <https://doi.org/10.1145/3597638.3608382>
- [135] José Morais, Luz Cary, Jesús Alegria, and Paul Bertelson. 1979. Does awareness of speech as a sequence of phones arise spontaneously? *Cognition* 7, 4 (1979), 323–331.
- [136] Michael Mulshine, Ge Wang, Chris Chafe, Jack Atherton, terry feng, and Celeste Betancur. 2023. WebChuckK: Computer Music Programming on the Web. , Article 28 (May 2023), 6 pages. http://nime.org/proceedings/2023/nime2023_28.pdf
- [137] Virginia Murphy-Berman and Linda Whobrey. 1983. The Impact of Captions On Hearing-Impaired Children's Affective Reactions To Television. *The Journal of Special Education* 17, 1 (April 1983), 47–62. <https://doi.org/10.1177/002246698301700107>
- [138] National Eye Institute 2019. *Types of color blindness*. National Eye Institute. <https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/color-blindness/types-color-blindness> Accessed on August 3, 2022.
- [139] Marina Nespov, Mohinish Shukla, and Jacques Mehler. 2011. *Stress-Timed vs. Syllable-Timed Languages*. John Wiley & Sons, Ltd, Oxford, UK, Chapter 48, 1–13. <https://doi.org/10.1002/9781444335262.wbctp0048>
arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781444335262.wbctp0048>
- [140] Stephen Nixon, Lisa Huang, Katja Schimmel, Rafał Buchner, and Cris R Hernández. 2023. Recursive Sans & Mono. <http://www.recursive.design/>

- [141] Robert M Ochshorn and Max Hawkins. 2015. Gentle: a robust yet lenient forced aligner built on Kaldi. <https://lowerquality.com/gentle/>
- [142] 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.
- [143] Desmond Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. 2019. Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset. *IEEE Transactions on Affective Computing* 12 (2019), 579–594.
- [144] Bryan Orme. 2009. *Maxdiff analysis: Simple counting, individual-level logit, and hb*. Technical Report. Sawtooth Software.
- [145] Hilary Palmén, Michael Gilbert, and David Crossland. 2023. How bold can we be? The impact of adjusting font grade on readability in light and dark polarities. 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 402, 11 pages. <https://doi.org/10.1145/3544548.3581552>
- [146] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (<conf-loc>, <city>San Francisco</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) (UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. <https://doi.org/10.1145/3586183.3606763>
- [147] Rosalind W. Picard. 1997. *Affective Computing*. The MIT Press. <https://doi.org/10.7551/mitpress/1140.001.0001>
- [148] Nita Prabhu, Luis Vargas, and Xiaogang Hu. 2022. Quantitative Characterization of Haptic Sensory Adaptation Evoked Through Transcutaneous Nerve Stimulation. In *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*. IEEE, Orlando, Florida, USA., 1–4. <https://doi.org/10.1109/ICHMS56717.2022.9980598>
- [149] Suksumek Promphan. 2017. Emotional Type: Emotional expression in text message.

- [150] 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](https://arxiv.org/abs/2212.04356)
- [151] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. <https://doi.org/10.48550/ARXIV.2212.04356>
- [152] Dhevi J Rajendran, Andrew T Duchowski, Pilar Orero, Juan Martínez, and Pablo Romero-Fresco. 2013. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives* 21, 1 (2013), 5–21.
- [153] Carlos Ramos-Carreño. 2022. dcor: distance correlation and energy statistics in Python. <https://doi.org/10.5281/zenodo.3468124>
- [154] 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>
- [155] Raisa Rashid, Jonathan Aitken, and Deborah I Fels. 2006. Expressing emotions using animated text captions. In *International Conference on Computers for Handicapped Persons*. Springer, Linz, Austria, 24–31.
- [156] 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.
- [157] Nancy A Remington, Leandre R Fabrigar, and Penny S Visser. 2000. Reexamining the circumplex model of affect. *Journal of personality and social psychology* 79, 2 (2000), 286.
- [158] Tara Rosenberger-Shankar and Ronald L MacNeil. 1999. Prosodic font: translating speech into graphics. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, Pittsburgh, USA, 252–253.
- [159] Jazz Rui Xia Ang, Ping Liu, Emma McDonnell, and Sarah Coppola. 2022. “In This Online Environment, We’re Limited”: Exploring Inclusive Video Conferencing Design for Signers. 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 609, 16 pages. <https://doi.org/10.1145/3491102.3517488>
- [160] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

- [161] Vít Rusňák, Pavel Troubil, Desana Daxnerová, Pavel Kajaba, Matej Minárik, Svatoslav Ondra, Tomáš Sklenák, and Eva Hladká. 2016. CoUnSiL: A video conferencing environment for interpretation of sign language in higher education. In *2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET)*. IEEE, Istanbul, Turkey, 1–8. <https://doi.org/10.1109/ITHET.2016.7760711>
- [162] Marie-Laure Ryan. 2007. Toward a definition of narrative. *The Cambridge companion to narrative* 22 (2007), 22–35.
- [163] Katri Salminen, Veikko Surakka, Jani Lylykangas, Jussi Rantala, Teemu Ahmaniemi, Roope Raisamo, Dari Trendafilov, and Johan Kildal. 2012. Tactile Modulation of Emotional Speech Samples. *Advances in Human-Computer Interaction* 2012 (2012), 1–13. <https://doi.org/10.1155/2012/741304>
- [164] Tim Schlippe, Shaimaa Alessai, Ghanimeh El-Taweel, Matthias Wölfel, and Wajdi Zaghouni. 2020. Visualizing Voice Characteristics with Type Design in Closed Captions for Arabic. In *2020 International Conference on Cyberworlds (CW)*. IEEE, Caen, France, 196–203.
- [165] Florian J. Schmidt-Skipiol and Peter Hecker. 2015. Tactile Feedback and Situation Awareness - Improving Adherence to an Envelope in Sidestick-Controlled Fly-by-Wire Aircrafts. In *15th AIAA Aviation Technology, Integration, and Operations Conference*. American Institute of Aeronautics and Astronautics. <https://doi.org/10.2514/6.2015-2905>
- [166] Mark Seidenberg. 2017. *Language at the Speed of Sight: How we Read, Why so Many Can't, and what can be done about it*. Basic Books.
- [167] Hasti Seifi and Karon E. MacLean. 2013. A first look at individuals' affective ratings of vibrations. In *2013 World Haptics Conference (WHC)*. 605–610. <https://doi.org/10.1109/WHC.2013.6548477>
- [168] Matthew Seita, Khaled Albusays, Sushant Kafle, Michael Stinson, and Matt Huenerfauth. 2018. Behavioral Changes in Speakers Who Are Automatically Captioned in Meetings with Deaf or Hard-of-Hearing Peers. 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, 68–80. <https://doi.org/10.1145/3234695.3236355>
- [169] Matthew Seita and Matt Huenerfauth. 2020. Deaf Individuals' Views on Speaking Behaviors of Hearing Peers When Using an Automatic Captioning App. 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, 1–8. <https://doi.org/10.1145/3334480.3383083>
- [170] 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>
- [171] Kristen Shinohara and Jacob O. Wobbrock. 2011. In the shadow of misperception: assistive technology use and social interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Vancouver</city>, <state>BC</state>, <country>Canada</country>, </conf-loc>) (*CHI '11*). Association for Computing Machinery, New York, NY, USA, 705–714. <https://doi.org/10.1145/1978942.1979044>
- [172] Chad Smith and Tamby Allman. 2019. Diversity in deafness: Assessing students who are deaf or hard of hearing. *Psychology in the Schools* 57, 3 (Oct. 2019), 362–374. <https://doi.org/10.1002/pits.22310>
- [173] Juho Snellman. 2015. Win probability? <https://github.com/sublee/trueskill/issues/1#issuecomment-149762508>
- [174] 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.
- [175] Freya Sukalla, Helena Bilandzic, Paul D. Bolls, and Rick W. Busselle. 2016. Embodiment of Narrative Engagement: Connecting Self-Reported Narrative Engagement to Psychophysiological Measures. *Journal of Media Psychology* 28, 4 (Oct. 2016), 175–186. <https://doi.org/10.1027/1864-1105/a000153>
- [176] Petra Sundström, Anna Ståhl, and Kristina Höök. 2007. In situ informants exploring an emotional mobile messaging system in their everyday practice. *International Journal of Human-Computer Studies* 65, 4 (April 2007), 388–403. <https://doi.org/10.1016/j.ijhcs.2006.11.013>

- [177] 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.
- [178] 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>
- [179] David Ternes and Karon E. MacLean. 2008. *Designing Large Sets of Haptic Icons with Rhythm*. Springer Berlin Heidelberg, 199–208. https://doi.org/10.1007/978-3-540-69057-3_24
- [180] Alexander Toet. 2022. *EmojiGrid*. Open Science Framework. <https://doi.org/10.17605/OSF.IO/H82YB>
- [181] Alexander Toet, Daisuke Kaneko, Shota Ushiana, 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>
- [182] 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>
- [183] 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>
- [184] Andreas Triantafyllopoulos. 2022. Personal communication.
- [185] Andreas Triantafyllopoulos, Johannes Wagner, Hagen Wierstorf, Maximilian Schmitt, Uwe Reichel, Florian Eyben, Felix Burkhardt, and Björn W. Schuller. 2022. Probing Speech Emotion Recognition Transformers for Linguistic Knowledge. <https://doi.org/10.48550/ARXIV.2204.00400>
- [186] Walda Verbaenen. 2019. *Phonotype. The visual identity of a language according to its phonology*. Master's thesis. PXL-MAD.
- [187] Ronald T. Verrillo. 1992. Vibration Sensation in Humans. *Music Perception* 9, 3 (04 1992), 281–302. <https://doi.org/10.2307/40285553>

- [188] Christian Vogler, Paula Tucker, and Norman Williams. 2013. Mixed Local and Remote Participation in Teleconferences from a Deaf and Hard of Hearing Perspective. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (Bellevue, Washington) (*ASSETS '13*). Association for Computing Machinery, New York, NY, USA, Article 30, 5 pages. <https://doi.org/10.1145/2513383.2517035>
- [189] Celina Isabelle von Eiff, Sascha Frühholz, Daniela Korth, Orlando Guntinas-Lichius, and Stefan Robert Schweinberger. 2022. Crossmodal Benefits to Vocal Emotion Perception in Cochlear Implant Users. *iScience* 25, 12 (2022), 105711.
- [190] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2022. Dawn of the transformer era in speech emotion recognition: closing the valence gap. <https://doi.org/10.48550/ARXIV.2203.07378>
- [191] 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>
- [192] 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) (*CHIEA '20*). Association for Computing Machinery, New York, NY, USA, 1–9. <https://doi.org/10.1145/3334480.3382985>
- [193] 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>
- [194] Ge Wang, Perry R. Cook, and Spencer Salazar. 2015. Chuck: A Strongly Timed Computer Music Language. *Computer Music Journal* 39, 4 (12 2015), 10–29. https://doi.org/10.1162/COMJ_a_00324 arXiv:https://direct.mit.edu/comj/article-pdf/39/4/10/1953737/comj_a_00324.pdf
- [195] Jianji Wang and Nanning Zheng. 2020. Measures of Correlation for Multiple Variables. arXiv:1401.4827 [math.ST]

- [196] 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>
- [197] Janet M. Weisenberger, Susan M. Broadstone, and Frank A. Saunders. 1989. Evaluation of two multichannel tactile aids for the hearing impaired. *The Journal of the Acoustical Society of America* 86, 5 (Nov. 1989), 1764–1775. <https://doi.org/10.1121/1.398608>
- [198] John D. Wells, Damon E. Campbell, Joseph S. Valacich, and Mauricio Featherman. 2010. The Effect of Perceived Novelty on the Adoption of Information Technology Innovations: A Risk/Reward Perspective. *Decision Sciences* 41, 4 (2010), 813–843. <https://doi.org/10.1111/j.1540-5915.2010.00292.x> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5915.2010.00292.x>
- [199] 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>
- [200] Matthew Wickline. 2001. Coblis – Color blindness simulator. <https://www.color-blindness.com/coblis-color-blindness-simulator/>
- [201] David Wicks. 2017. The coding manual for qualitative researchers. *Qualitative research in organizations and management: an international journal* 12, 2 (2017), 169–170.
- [202] Deirdre Wilson and Tim Wharton. 2006. Relevance and Prosody. *Journal of Pragmatics* 38, 10 (Oct. 2006), 1559–1579. <https://doi.org/10.1016/j.pragma.2005.04.012>
- [203] Werner Wirth, Tilo Hartmann, Saskia Böcking, Peter Vorderer, Christoph Klimmt, Holger Schramm, Timo Saari, Jari Laarni, Niklas Ravaja, Feliz Ribeiro Gouveia, Frank Biocca, Ana Sacau, Lutz Jäncke, Thomas Baumgartner, and Petra Jäncke. 2007. A Process Model of the Formation of Spatial Presence Experiences. *Media Psychology* 9, 3 (May 2007), 493–525. <https://doi.org/10.1080/15213260701283079>
- [204] John Wiseman. 2021. py-webtrcvad. <https://github.com/wiseman/py-webtrcvad>.

- [205] Matthias Wölfel, Tim Schlippe, and Angelo Stitz. 2015. Voice driven type design. In *2015 international conference on speech technology and human-computer dialogue (SpeD)*. IEEE, Bucharest, Romania, 1–9.
- [206] Lei Zhang and Doug A. Bowman. 2022. Exploring Effect of Level of Storytelling Richness on Science Learning in Interactive and Immersive Virtual Reality. In *Proceedings of the 2022 ACM International Conference on Interactive Media Experiences (Aveiro, JB, Portugal) (IMX '22)*. Association for Computing Machinery, New York, NY, USA, 19–32. <https://doi.org/10.1145/3505284.3529960>
- [207] Nianmei Zhou, Steven Devleminck, and Luc Geurts. 2024. Tangible Affect: A Literature Review of Tangible Interactive Systems Addressing Human Core Affect, Emotions and Moods. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (Copenhagen, Denmark) (DIS '24)*. Association for Computing Machinery, New York, NY, USA, 424–440. <https://doi.org/10.1145/3643834.3661608>

Appendices

APPENDIX A

Interview Protocol for Study 1

A.1 *Introduction*

A.1.1 *Brief Greeting*

Thanks to the participant and introduce yourself (Experimenter).

A.1.2 *Goal/Purpose of the Study*

This is a study about how online meetings that use automatic captions between hearing and Deaf and Hard-of-Hearing individuals can sometimes fail, especially from the point of view of DHH persons. We want to understand in what situations they can go wrong, and what happens when they do.

A.1.3 *Procedure*

This interview will be divided into two parts. First, we will ask you a bit about your experience with online meetings, whether for work, school, or your personal life. Following, we will show you three videos of design ideas we think could be added to online meeting software, and will ask you your thoughts about them.

A.2 *The Interview*

A.2.1 *Participants' Experiences*

1. Where do you typically use online meeting software?
2. Was there ever a situation where you felt left out of a conversation that was going between hearing participants in a meeting? How was it?
 - 2.a. How often do things like this happen?
 - 2.b. How do you deal with it? Do you have any strategies to make it better?
3. Do you ever feel as if things hearing individuals are saying are hard to understand?
 - 3.a. Can you give us some examples of this happening?
 - 3.b. Can you elaborate why you think that is the case?
 - 3.c. Would you say that some speakers are harder to understand than others?
 - 3.d. What makes them more difficult to understand than others?
 - 3.e. What does an easy-to-understand speaker do that sets them apart?
4. Even when words are correctly represented, do you ever feel as if something the speaker said is missing when it is presented to you through an interpreter or captions?
 - 4.a. Can you give us some examples?
 - 4.b. Can you elaborate on why you think this happens?
 - 4.c. If something is missing, what do you think it is?
 - 4.d. Have you ever wondered how these 'missing' elements could be made available?
5. When a hearing person is speaking, how do you usually figure out their mood or emotions?
6. Do you use any accessibility tools, such as live captions, multi-pinning, etc.?
 - 6.a. If you've ever used live captions, what is your impression of them?
 - 6.b. What do you think they do well?
 - 6.c. Where do you think they could improve?
 - 6.d. When captions are correct, do you ever feel as if something is still missing? What?

A.2.2 *New Ideas for Meeting Software*

We are now going to discuss a few ideas we are working on to enhance online meeting software. We are interested in exploring ways of representing elements of the voice that are not represented in traditional automatic captions.

1. What do you think those elements could be?
 - 1.a. Have you ever felt that this loss was consequential? How?
 - 1.b. If you were to make these missing elements accessible for DHH individuals, how would you do it?
2. Have you ever felt valence, or emotion in general, was something useful in speech but missing in traditional meeting software?
3. Do you remember any situation where this lack of emotion in captions was an issue for you?
4. How do you think valence could be presented to DHH individuals?
5. Have you ever felt prosody was useful in speech but missing in traditional meeting software?
6. Do you remember any situation where this lack of prosody in captions was an issue for you?
7. How do you think prosody could be presented to DHH individuals?
8. Between valence and prosody, which do you think is more useful? Why?

A.3 *Prototype Demo and Preliminary Evaluation*

We are now going to show you three videos where a speaker's voice has some of these elements overlaid on the interface. Imagine that you were talking with this person on the video in an online meeting software, and that they are telling you a story of their life. Each video explores a different interface element.

(The three demo videos were shown.)

1. Which of the three prototypes did you like the most? Why?

2. Which did you like the least? Why?
3. Considering each prototype:
 - 3.a. Border changing color and thickness – what worked well, what didn't?
 - 3.b. Words changing color and thickness – what worked well, what didn't?
 - 3.c. Font shape changing – what worked well, what didn't?
4. In general, what do you feel the design ideas added to what was being said?
5. Do you see yourself using these features? In what situations?
6. Do you have any suggestions on how you would do things differently?
7. Any last comments?

A.4 *Collect Demographic Information*

A.5 *Exit the Interview*

Thank the participant for their time and close the session.

APPENDIX B

12-Item Narrative Engagement Scale

The questions below, adapted from Busselle and Bilandzic [35], were administered to participants after each of the five videos in Study 2. Although grouped here by their four subscales, the experiment randomized their order for each participant, who was unaware of these groupings.

1. *Narrative Understanding*

- (a) At points, I had a hard time making sense of what was going on in the video.
- (b) My understanding of the characters is unclear.
- (c) I had a hard time recognizing the thread of the story.

2. *Attentional Focus*

- (a) I found my mind wandering while the video was on.
- (b) While the video was on I found myself thinking about other things.
- (c) I had a hard time keeping my mind on the video.

3. *Narrative Presence*

- (a) During the video, my body was in the room, but my mind was inside the world created by the story.
- (b) The video created a new world, and then that world suddenly disappeared when the video ended.
- (c) At times during the video, the story world was closer to me than the real world.

4. *Emotional Engagement*

- (a) The story affected me emotionally.
- (b) During the video, when the speaker was happy, I felt happy, and when they suffered in some way, I felt sad.
- (c) I felt sorry for the speaker in the video.

*This dissertation's main text was set in John Hudson's Brill typeface,
with Rasmus Andersson's Inter used as its sans serif companion
for the numbering of pages, parts, sections, and more.*