Tactile Emotions: Multimodal Affective Captioning with Haptics Improves Narrative Engagement for d/Deaf and Hard-of-Hearing Viewers

CALUÃ DE LACERDA PATACA, Comp. & Info. Sciences, Rochester Institute of Technology, USA
SAAD HASSAN, Department of Computer Science, Tulane University, USA
LLOYD MAY, Music Department, Stanford University, USA
MICHELLE M OLSON, School of Information, Rochester Institute of Technology, USA
TONI D'AURIO, Department of ASL and Interpreting Education, Rochester Institute of Technology, USA
ROSHAN L PEIRIS, School of Information, Rochester Institute of Technology, USA
MATT HUENERFAUTH, School of Information, Rochester Institute of Technology, USA

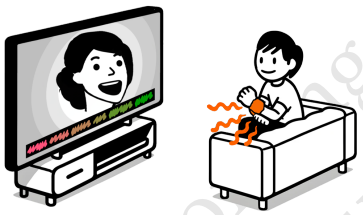


Fig. 1. Multimodal affective captions, combining visual cues and vibrations felt via a wrist-worn device, enrich the viewing experience for d/Deaf or Hard-of-Hearing individuals by portraying speaker emotions, improving engagement.

This paper explores a multimodal approach for translating emotional cues present in speech, designed with Deaf and Hard-of-Hearing (DHH) individuals in mind. Prior work has focused on visual cues applied to captions, successfully conveying whether a speaker's words have a negative or positive tone (valence), but with mixed results regarding the intensity (arousal) of these emotions. We propose a novel method using haptic feedback to communicate a speaker's arousal levels through vibrations on a wrist-worn device. In a formative study with 16 DHH participants, we tested six haptic patterns and found that participants preferred single per-word vibrations at 75 Hz to encode arousal. In a follow-up study with 27 DHH participants, this pattern was paired with visual cues, and narrative engagement with audio-visual content was measured. Results indicate that combining haptics with visuals significantly increased engagement compared to a conventional captioning baseline and a visuals-only affective captioning style.

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

Additional Key Words and Phrases: Accessibility, Emotion / Affective Computing, Individuals with Disabilities & Assistive Technologies, Empirical study that tells us about how people use a system

Authors' Contact Information: Caluã de Lacerda Pataca, Comp. & Info. Sciences, Rochester Institute of Technology, Rochester, NY, USA, calua.pataca@gmail.com; Saad Hassan, Department of Computer Science, Tulane University, New Orleans, LA, USA, saadhassan@tulane.edu; Lloyd May, Music Department, Stanford University, Palo Alto, CA, USA, lloydmay@stanford.edu; Michelle M Olson, School of Information, Rochester Institute of Technology, Rochester, NY, USA, mmo9420@rit.edu; Toni D'Aurio, Department of ASL and Interpreting Education, Rochester Institute of Technology, Rochester, NY, USA, tld3799@rit.edu; Roshan L Peiris, School of Information, Rochester Institute of Technology, Rochester, NY, USA, rxpics@rit.edu; Matt Huenerfauth, School of Information, Rochester Institute of Technology, Rochester, NY, USA, matt.huenerfauth@rit.edu.

1 Introduction

Speech-modulated typography has seen increased interest by authors in the fields of Human-Computer Interaction and Computing Accessibility. Researchers have explored varied approaches to making the visual appearance of written words echo the expressive qualities of the speaker who said them [15, 23, 24, 39–41, 49, 71, 75, 90]. For example, a word spoken loudly could be written in a larger font; the words of a sad speaker might appear in thin, pale-red letters; song lyrics could be displayed as if they were notes on a musical score, and so on.

One important application for these approaches is that of improving the accessibility of speech for d/Deaf and Hard-of-Hearing (DHH) people. Leveraging advancements in automatic speech recognition, captions have become a commoditized and ubiquitous feature in many digital systems that feature oral speech. Making them better is an understandable goal—they are everywhere!—and the fact that their technological underpinnings now provides accurate digital pathways to translate spoken sound into written speech opens up interesting and novel possibilities for its achievement.

There is nuance, though, in capturing and depicting non-speech information through captions. Speech is a rich signal, and research has shown that presenting its words along with some of its features (but not others) can lead to improved understanding for DHH viewers. Namely, authors found that DHH users preferred text that conveyed a speaker's emotions over text depicting the actual sound of their voices. In other words, *affective* captions were preferred over captions based on acoustic features [25].

This finding has led researchers to explore the design space of *affective captions* [24, 41, 49]. Here, visual cues applied to each word are used to indicate the emotions in speakers' voices, which can range from positive to negative (valence) and from calm to excited (arousal). Both dimensions are important, but while researchers found that DHH participants clearly preferred color-coding for indicating the speaker's valence, no single visual modulation proved equally effective for conveying their arousal levels [24]. Thus, even though using visual cues to convey emotional valence seems appropriate, we hypothesize that the depiction of arousal might benefit from a distinct approach. To this end, in this paper *we propose and evaluate the use of haptic feedback in affective captions to represent a speaker's arousal levels.*

In so doing, our work is inspired by Akshita et al., who found that in multi-modal visual-haptic stimuli, the visual aspect shapes participants' perception of valence, while the haptic component influences their interpretation of arousal [2]. We have adapted this concept to the realm of affective captions: In a first, formative study, we sought to understand DHH viewers' preferences towards different haptic patterns used to convey arousal. With a preferred pattern identified, we next investigated its impact on viewers' *narrative engagement* with audio-visual content, i.e., how the haptic feedback influenced their emotional connection and overall immersion in the story.

As such, we offer three main contributions:

(1) We propose a novel approach to encode arousal levels inferred from speech as haptic feedback conveyed to users through a wrist-worn device. This can be combined with visual *affective captions,* as has been explored in prior work;
(2) An assessment, in Study 1, of six distinct haptic patterns to determine their effectiveness in conveying the arousal levels of a speaker to DHH viewers. Our analysis revealed a preferred pattern among the majority of participants: a single, short pulse per word, vibrating at 75 Hz and with its amplitude varying according to the intensity levels of arousal;
(3) An investigation, in Study 2, of how affective captioning strategies influence DHH viewers' *narrative engagement.* We combined the preferred haptic pattern from Study 1 with visual representations of valence and/or arousal and compared them them to a neutral baseline.

Our findings showed that the combination of haptics plus visual cues significantly outperformed both the baseline and a visuals-only affective caption style.

2 Background and related work

In this section, we review prior work on the use of typographic modulations to depict paralinguistic dimensions of speech, including prosody, manner of speech, and affective states. Some of this research has shown that while valence can be effectively conveyed through purely visual means, there is potential to depict arousal through non-visual channels [24]. This is a driving force behind our present work, i.e., investigating the representation of speech arousal using both visual and haptic channels. Additionally, we will go over studies on how haptics can enhance the perception of speech and music for DHH individuals, and their potential for conveying emotions, which informs our approach.

2.1 Using typography to convey changes in a speaker's tone of voice

While spoken and written language are connected, they are not perfect mirrors of each other [73]. Writing systems are shaped by unique constraints, including the need to reduce writing effort. As a result, many elements present in speech are omitted in written text, as readers are often expected to infer them from context [73]. However, this can lead to confusion, as a sentence might have multiple meanings. Unlike spoken language, where a speaker can clarify their intended meaning through changes in tone of voice [88], a reader might struggle with the ambiguity of text.

Consider, for example, the sentence "I didn't say you are funny." Different meanings emerge depending on which word a speaker emphasizes. *"I* didn't say..." suggests that someone else may have said it; "I didn't *say* you..." implies that *you are funny* was conveyed, just not verbally; "...you are *funny*" hints that something was indeed said, but it wasn't the word *funny.* Subtle shifts in meaning like these highlight challenges that readers may face when interpreting written text that is presented without the disambiguating aid of vocal tone.

This gap between spoken language and written text has been the focus of different authors, who have explored ways to bridge it. Some have developed tools allowing writers to embed cues in their text, guiding readers toward an intended pronunciation or tone. For example, Verbaenen [80] modified the Times New Roman typeface to visually distinguish between similarly sounding phonemes in Dutch. Similarly, Bessemans et al. [6] worked with visual changes to letter shapes to convey elements of prosody—such as pitch, rhythm, and loudness—to help novice readers improve their expressive reading skills.

When we consider automatic captioning systems, the traditional concept of a "writer" may not strictly apply. Similarly to manually generated captions, these systems translate auditory information into a visual format, making spoken language accessible to readers; However, their structure reflects the design choices of the originating system's developers rather than the intentions of the speaker or the interpretations of a human captioner. Typically with these systems, non-verbal cues—such as tone, rhythm, and emotion—are omitted, which, as studies have shown, can lead to communication breakdowns. This is particularly true for DHH individuals who depend on these systems to access both the explicit content and the subtler nuances of spoken communication [25, 52].

To reduce this disconnect between voice and its (automated) transcription, several approaches have been explored. Some authors [15, 23, 68, 71, 90] have worked on mapping acoustic features from the speech signal to visual parameters of its textual transcriptions. For instance, de Lacerda Pataca and Costa mapped loudness to font-weight and pitch to baseline-shift [22], with a follow-up study adding rhythm mapped to letter-spacing [23]. In this approach, a passage that is slow, high-pitched, and quiet would be displayed in widely spaced, vertically raised, and thin letters. This could allow readers to imagine the sound of the spoken utterance when reading its

speech-modulated transcription, thereby inferring how its sound qualities influenced its meaning [22, 23].

Recent studies suggest that in accessibility applications targeting DHH users, the choice of speech features to include matters. While useful, depicting acoustic features can significantly reduce legibility [25]. As an alternative, some researchers propose displaying *affective* cues in typography [25, 41, 49, 75]. This approach focuses on representing emotional dimensions, namely, valence and arousal, rather than raw acoustic features. According to Russell's circumplex model [69], valence refers to the positivity or negativity of an emotion, mapped to an x-axis, while arousal refers to the level of excitement or calmness, mapped to a y-axis. Using machine-learning models [82], these features can be derived from speech signals and then mapped to visual cues in text.

Exploring the design space of affective captions from the point of view of DHH individuals, de Lacerda Pataca et al. [24] found evidence that font color, modulated within a color scale defined by Hassan et al. [41], effectively conveyed differences in valence. The representation of arousal, however, was less straightforward. Although font-weight and font-size were suggested for depicting arousal, preferences for these styles varied widely [24].

These trends in previous research suggest that merely altering the visual appearance of typography may be insufficient to effectively communicate a speaker's arousal levels, particularly without additional reinforcing cues. Since arousal is a key dimension of the circumplex model, it is important to explore alternative methods. Haptic feedback, as we will discuss in the following section, has previously been used both as a complement to captions and as a method for conveying emotions, making it a promising candidate for complementing visuals-only affective captions.

2.2 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 [56]. These technologies have a wide range of applications, from tactile feedback systems that help pilots maintain safe flight parameters [72] to enhancing medical training [3]. Within the HCI and accessibility communities, there has been growing interest in using haptics to convey aspects of sound, such as speech [32], environmental sounds [45], and music [31, 58], 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 [29]. Because of their mainstream appeal, they can avoid the stigma often associated with dedicated assistive devices [76]. 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. [36].

Other haptic devices have also been used to complement visual information. Weisenberger et al. [86], for example, found that translating sound into tactile signals improved speech reading accuracy for DHH people. This was echoed by Fletcher et al. [32], who showed that haptic feedback can enhance speech intelligibility for cochlear implant users, particularly after a period of training.

In another study, Wang et al. [85] explored translating speech sounds into haptic feedback, helping their DHH participants differentiate between speakers and intuit their moods. Remarkably, they achieved this using a simple setup: a voice coil actuator placed in a 3D-printed wrist-worn casing, driven by a 3W power amplifier—cheap and readily available components. For our studies, we employed a similar setup (see subsection 3.2 for more details).

In essence, these methods exemplify the concept of *sensory substitution*, where one sense is supplemented with information that would typically be gathered by another [57]. 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 [33]. 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 [30] compared to touch at different parts of the skin [18, 81]. Privacy and comfort concerns also arise, particularly when music or human speech is used as direct input to vibrotactile haptic systems. Such signals contain frequencies within 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 [59]. Verrillo [81], for instance, found that the *sensational quality* of vibrations below 100 Hz differs 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 [79], 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 [74], who found that participants assigned different affective categories to different stimuli—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 our first study, detailed in subsection 4.1.

Akshita et al. [2] 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.

2.3 Research questions

As we have seen, affective captions have the potential of becoming an important approach to making speech more accessible, in particular for DHH people [25]. Despite advancements, research is still needed to improve on their form which—we speculate—might gain from a haptic-based complement. Since this is still an as of yet unexplored space, our first study hopes to answer *how* this haptic signal could be shaped. From among a set of rhythmic patterns and frequencies suggested in the literature, we ask:

RQ1 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?

In this formative study, we imagine that the haptic signal's amplitude will be modulated to convey different arousal levels (e.g., high arousal → strong vibrations, low arousal → weak vibrations) with the goal being to find the rhythmic pattern and frequency that best accommodate these changes.

Following the findings from this first formative study, we investigate the overall effectiveness of combining haptic feedback with visual modulations in representing arousal states. Specifically, we aim to understand how these modalities interact and whether their combined effect enhances the perception of affective states in speech, ultimately leading to improved *narrative engagement*[1]:

RQ2 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?

[1]This construct measures dimensions of engagement, including empathetic response and immersion. A detailed definition is provided in section 5.1.3.

3 System Design

This section describes the design of the haptic-captioning system used in Studies 1 and 2. Although there were differences in the setup for each study, both employed the same core functionality, which is presented in full here. First, we present our pipeline to process each video's audio files, obtaining speech transcriptions with corresponding affective features (3.1). 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 (3.2). Third, we discuss how we implemented the visuals applied to the typography of the captions used in the two studies (3.3).

3.1 Transcription and Emotion Recognition of a Speech Signal

All videos were transcribed using OpenAI's Whisper speech recognition model [67] with word-level timestamping [55]. The voice activity detection (VAD) flag was enabled to improve transcription when background noises were present.

We employed the circumplex *dimensional* model of emotions [69]. In it, emotions are represented as coordinates on a plane that maps their position along unpleasant-pleasant (valence) and calm-excited (arousal) axes. Thus, where a categorical model might use a discrete label to define an emotion as *sad*, the circumplex model characterizes it by low valence and arousal levels.

While there are compelling examples of affective captions and typography that employ categorical models (e.g., [44, 54]), we follow de Lacerda Pataca et al. [25], who argued that dimensional models like Russell's circumplex model allow a viewer to consider emotional states conveyed in speech in a more nuanced and less prescriptive manner. The way that different emotions can overlap in their valence and arousal values, they argue, allows for greater interpretive flexibility, which in turn allows viewers to better integrate other contextual cues such as the overall story arc, facial expressions, body language, etc, contributing to a richer and more contextually-grounded understanding of emotional content.

To implement this approach, we deployed Wagner et al.'s open-source toolkit [82] 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 [21].

3.2 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 this paper's contribution is our novel approach to do so, i.e., the way we modulate the intensity of perceived vibrations 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 [79], 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. [2], comprise a haptic signal: frequency and rhythm.[2]

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 [29], we followed Wang et al.'s haptic captioning study [85] and used Acouve's Vp2 Vibro-Transducer,[3] a simple voice coil driven

[2] Akshita 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 [2].
[3] https://www.acouve-lab.com/products

Fig. 2. 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.

by Techtile Toolkit's power amplifier [60]. 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 2.

To generate the audio files driving the haptic patterns, we wrote a ChucK language script [84] that converted arousal values encoded in the caption file into a sound signal to be played alongside the video. ChucK 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.[4] 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 then used to route these outputs to their respective devices.

3.3 Typographic Representations of Valence and Arousal Levels

Both studies used typographic modulations to convey speech features. Study 1 focused solely on valence, while Study 2 explored different combinations of valence, arousal, or neither. To modulate captions with these values, we based our approach on prior work [24, 25, 41, 49].

Using the per-word emotion levels obtained through our speech-analysis model, our system was able to render captions where each word's visual style is changed to reflect changes in valence and arousal. The speech → typography mappings we used are based on prior work that systematically evaluated competing typographic styles with DHH participants, aiming to find a combination that

[4]The actual implementation was done using the WebChucK toolkit [61], which can run in a web environment and, as such, could be integrated into the same web-based script, described in de Lacerda Pataca [21], we used to generate the affective captions.

Fig. 3. 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.'

was both highly preferred and effective in conveying the speaker's emotional cues [24]. The design recommendation we followed suggests depicting valence through changes in font-color, and arousal through changes in font-weight.

For the color scale applied to valence, we followed Hassan et al.'s orange-red (for negative) to white (for neutral) to aqua (for positive), since it is relatively resilient to less severe degrees of color-vision deficiencies [41]. To represent changes in arousal through variations in font-weight, we used the *Recursive* typeface by ArrowType foundry [62]. This is a variable font[5] that offers a wide-ranging font-weight axis, spanning from light (300) to extra-black (1000). An example of this font color / font weight modulation is shown in Figure 3.

4 Study 1: Formative Exploration of Haptic Patterns to Convey Speaker Arousal

Prior work established *intensity* as the primary haptic dimension for communicating differences in arousal. Study 1 aims to experimentally determine which *frequency* and *rhythmic* properties of this signal are perceived as effective and comfortable for doing so.

4.1 Defining the Different Haptic Patterns

4.1.1 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 rhythmic patterns from the set defined by Ternes and MacLean [79]. These patterns are listed below and schematically represented in Figure 4:

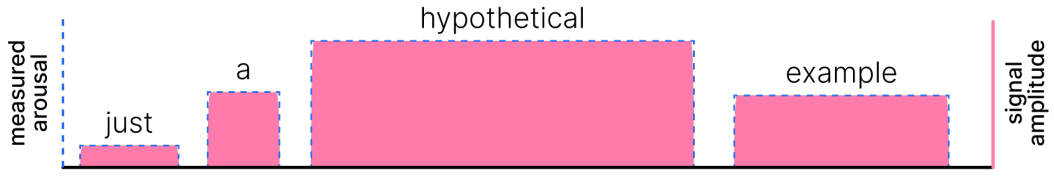
- LP A LONG PULSE, vibrating for the whole duration[6] of the word (Figure 4a);
- SSP A SINGLE SHORT PULSE lasting for two-thirds of the duration of the word, with a one-third silence at its end (Figure 4b);
- MSP A series of MULTIPLE SHORT PULSES with a fixed duration.[7] The number of pulses will be proportional to the duration of the word itself (Figure 4c).

4.1.2 Frequency. Along with *rhythm*, Akshita et al. [2] conceptualizes *frequency* as a defining property of a haptic signal. In essence, this is the rate at which the haptic signal oscillates, typically

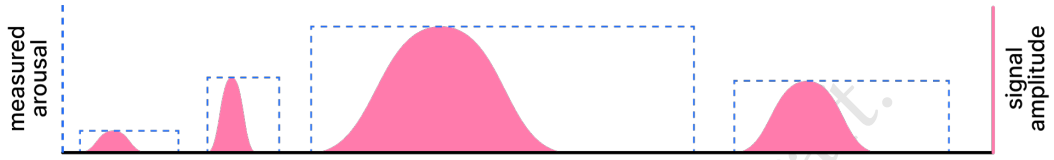
[5]In a *variable font* [19], the position of each point defining a glyph's visual contours can shift along different *axes* of variation. These variations are independent of one another, with the resulting glyph being derived as an interpolation of all these shifts applied to each variation axis. The Recursive font, for instance, includes axes that modulate its font-weight, slant, width, cursiveness, and "informality."

[6]To 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.

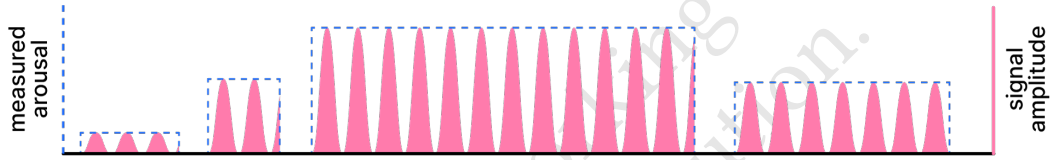
[7]The duration of these pulses follows Seifi and MacLean [74], who defined their fast-pulses rhytmic pattern as lasting 1/16th of a second each.



(a) LP: One continuous vibration per word.



(b) SSP: Vibration intensity increases for 1/3 of the word duration, then decreases for another 1/3, followed by 1/3 of silence.



(c) MSP: Pulse duration remains constant. Long words have many pulses, while short words have few.

Fig. 4. These charts illustrate how three haptic signal configurations (y-axis, right) respond to changing arousal levels (y-axis, left) over time (x-axis). 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."

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 [1], we defined two frequency levels: a low tier, at 75 Hz, and a high tier, at 250 Hz.

From previous perception literature [1], 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 75 Hz. 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 the authors 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 75 Hz waveform was approximately 4 dB lower than the 250 Hz waveform. Therefore, an offset of +6 dB was applied to the 250 Hz signal to approximately account for the frequency response effects of the hardware and for perceptual differences in an attempt to control for intensity as a confounding factor in the experiment.

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 1.

FREQUENCY	RHYTHMIC PATTERN		
	Long Pulse	*Single Short Pulse*	*Multiple Short Pulses*
75 Hz	$LP_{75\,Hz}$	$SSP_{75\,Hz}$	$MSP_{75\,Hz}$
250 Hz	$LP_{250\,Hz}$	$SSP_{250\,Hz}$	$MSP_{250\,Hz}$

Table 1. The six haptic conditions evaluated in Study 1.

4.2 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 d/Deaf or Hard-of-Hearing. For Study 1 we recruited a total of 16 participants, 9 of which identified as female and 7 as male, 11 of which identified as d/Deaf and 5 as Hard-of-Hearing, with a mean age of 27.1 years (σ = 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 [26]. The number of stimuli shown for each participant echoed studies [24] that employed a similar best-worst scaling setup (described below), i.e., 10 rounds, each with a different video, with four conditions tested per round. We randomized video and condition order, together with condition applied to each video.

The videos were sourced from the Stanford Emotional Narratives Dataset [63]. These are a set of unscripted, self-paced videos where a diverse group of people recount stories from their lives that have strong emotional overtones. While the dataset was originally created to aid the development of time-series emotion recognition models, its videos are useful for perceptual tests such as ours because of how short, emotionally rich, and visually homogeneous they are—all have a well lit and framed speaker sitting against a neutral backdrop—meaning, they allow us to show participants a large set of formally consistent stimuli in a relatively short session.

Short video excerpts were selected, and participants were asked to watch each one 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?"

4.2.1 Analysis plan. Building upon prior work looking at preferences regarding caption and typographic parameters [5, 24, 83], some of which targeted DHH individuals [5, 24], we adopted a best-worst scaling (BWS) methodology. This allows 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.

BWS is similar to pairwise comparisons in asking participants simple, scale-less "better or worse?" types of question. It has the advantage of leveraging *implied* answers to increase the number of data points collected in each round. For example, if participants are shown four conditions (*A*, *B*, *C*, and *D*) and explicitly identify pattern *A* as the best and *D* as the worst option, they are also implicitly indicating that *A* surpasses both *B* and *C*, and that *D* is worse than *B* and *C*. Thus, although participants only explicitly ranked two items in each round, we end up with five comparisons.

The method also offers some advantages over alternatives like integer rankings or Likert scales. Notably, by eschewing numeric scales, BWS mitigates inconsistencies in rating assessments [20, 50], making it particularly suitable for contexts where differences between conditions may be subtle [5].

Having the preference data collected as a set of pairwise comparisons allows us to employ an ELO-rating system in its analysis. In this system, each condition has a "rating," with which we can estimate how likely a participant would be to choose it over another option. As we initialize the analysis, all conditions start with the same rating, but as we process each comparison, these ratings are adjusted, taking into account who won and who lost at each step and, as more data points are included, the quality of predictions improves. Notably, the system is self-correcting, meaning, since a higher-rated condition's victory confirms the current ranking, when it happens it causes only a minimal adjustment to the rankings of both the winner's and the loser's rankings. Inversely, an upset victory by a lower-rated condition will lead to significant changes in the scores [27].

Following recent examples in HCI research [24, 65], we adopted Herbrich et al.'s TrueSkill implementation of an ELO-rating system [42].[8] The reasoning is that, beyond quantifying each condition's ranking, TrueSkill also provides estimates to the level of uncertainty around that value.

4.3 Findings from Study 1

4.3.1 Haptic pattern rankings. Study 1 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 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 related 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 5. Note that before processing participants' preferences, each haptic pattern was initialized with a skill of 25.[9]

4.3.2 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

[8]To address the order-dependency inherent in ELO-like systems, we supplemented TrueSkill with Clark et al.'s recommendation to average rankings across randomly ordered iterations until values stabilize.

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

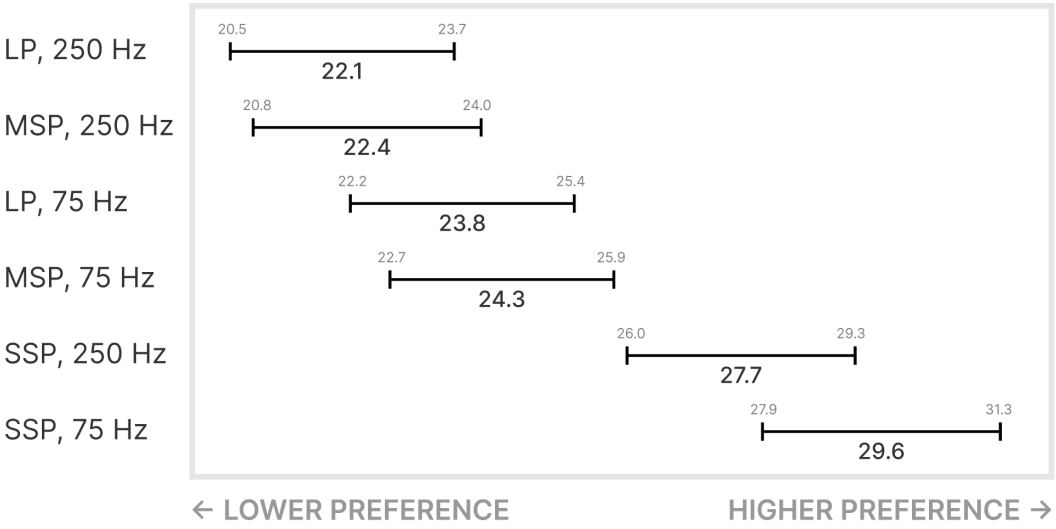


Fig. 5. Final TrueSkill rankings of the six haptic patterns. LP represents the long pulse; SSP, the single short pulse; MSP, the multiple short pulses. These are combined with two frequencies, 75 Hz and 250 Hz. The skill is shown below each line, with its 95% confidence range shown above.

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

Table 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 5. LP represents the long pulse; SSP, the single short pulse; MSP, the multiple short pulses.

the text seemed bland, unemotional, or matter-of-fact. Or calm, weak vibrations at a particularly emotional moment." Other participants echoed this sentiment, 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."

4.4 Discussion of Study 1

In response to RQ1, 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 5) 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 [78]. 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 times.

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 longer-form videos, ensuring participant comfort during the test was a key consideration.

While this study focused on the subjective preferences of participants, 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 hope to further explore

CONDITION	AROUSAL DEPICTION	VALENCE DEPICTION
$C1_B$ *(baseline)*	N / A	N / A
$C2_V$	VISUALS	VISUALS
$C3_H$	HAPTICS	VISUALS
$C4_{V+H}$	VISUALS & HAPTICS	VISUALS
$C5_N$	N / A	VISUALS

Table 3. The five conditions presented to participants in the second 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 $C5_N$ uses neither, showing only valence.

these themes in Study 2, with a primary focus on identifying the most engaging combination of haptic feedback and typographic modulations.

5 Study 2

Having established the SSP rhythmic pattern combined with the low frequency setting (75 Hz), we set out to answer our second 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.

5.1 Methods

5.1.1 Conditions. To answer RQ2, we defined four conditions regarding the portrayal of arousal:[10] arousal through visuals-only ($C2_V$), haptics-only ($C3_H$), visual *and* haptics ($C4_{V+H}$), and no arousal depiction ($C5_N$). We also included a conventional condition that had neither arousal nor valence as a baseline ($C1_B$). Table 3 summarizes the five conditions.

5.1.2 Stimuli. In selecting videos for Study 2, we followed three basic criteria:

(1) The videos should predominantly feature one speaker.[11]
(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 1, the videos 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 1 we sliced the videos to include these bursts, this approach would not have been suitable for the goals of Study 2. Here, measuring changes in narrative engagement required arousal levels that varied over a longer duration, meaning that a complete narrative arc needed to be established.

[10] Since our focus is not on the depiction of valence, in all of these conditions valence is shown through visuals (font-color).
[11] This criterion is based on the scrolling-caption-based style from prior literature (e.g., [24]), which has not yet been adapted or evaluated for multi-speaker settings—a topic outside the scope of our study.

NAME	SOURCE	ORIGINAL SOURCE
Sheriff Hassan's Monologue	*Midnight Mass,* episode 6, season 1	youtu.be/olhpqJso41M
Sally's Monologue	*Barry,* episode 7, season 2	youtu.be/qw62N4v8Cwo
The Arrival	Short film by Daniel Montanarini	vimeo.com/166075559
Scene from *Damage*	Short film by Matt Porter	vimeo.com/325243238
Scene from *The Human Voice*	Short film by Pedro Almodóvar	DVD copy

Table 4. The five videos used in the second study.

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).[12] From an initial pool of 26 titles, we processed 13 through the affective captioning pipeline (described in subsection 3.1). These were evaluated by two authors for narrative coherence and consistency between perceived and synthetically inferred emotions. Ultimately, five videos were selected for the study, as listed in Table 4. 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.

5.1.3 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 studies have primarily relied on two approaches: self-reported measures of usefulness [25, 49] and objective assessments of perceived valence and arousal levels [24, 41]. While these methods provide valuable insights, they also have limitations. Self-reported measures may be susceptible to novelty bias [87], 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, we propose using a different metric to capture the effects of affective captions, namely, *narrative engagement.*

Narrative engagement, as a measure, captures changes in cognitive processes in individuals as they attempt to make sense of a story [14]. It builds upon established constructs such as *spatial presence* (the sensation of being physically present and able to act within the depicted environment [53, 89]), *identification* (experiencing events portrayed in the narrative as if they were happening to oneself [17]), *flow / transportation* (becoming deeply absorbed in the narrative to the extent of losing self-awareness and awareness of surrounding events [13]), 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 [70].[13] As Gilbert suggests, human perception accepts mental representations as true, and disbelief requires additional cognitive steps [35]. 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.

[12] shortverse.com and shortoftheweek.com

[13] A 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 [13].

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 [7] (finding no significant differences), evaluating game narratives with low versus high fidelity graphics [11] (also finding no significant differences), and comparing automatic versus professionally authored closed captions for YouTube videos [48] (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.

5.1.4 Experimental design. Study 2 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.

5.1.5 Experimental procedure. Like with the first study, Study 2 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 randomized 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 A. 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 [14].

After finishing this section, they were presented with a screen, shown in Figure 6, 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 [12]. These were then discussed with other authors and refined.

5.2 Findings from Study 2

Recruitment, compensation, and inclusion criteria matched those of Study 1. 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 3. Where needed, participant quotes were edited for clarity and conciseness.

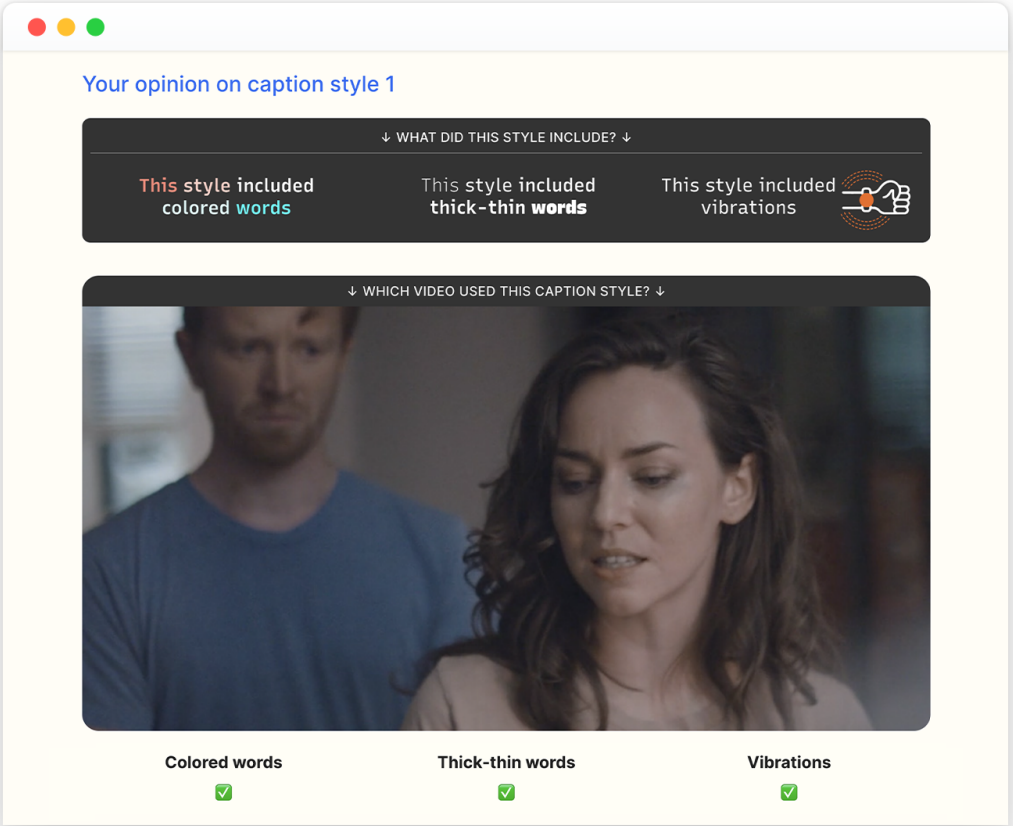


Fig. 6. Screenshot of the page where participants gave feedback about each condition. The image captured from the video was used as a mnemonic device for each caption condition, together with the illustrations and short descriptions. In this example, we see $C4_{V+H}$ (here labeled as "caption style 1"), which includes visuals and haptics for arousal, and visuals for valence.

5.2.1 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 5.

The distribution of scores is shown in Figure 7. Given that this is a within-subjects study with non-parametric data,[14] we used the Friedman test to compare the raw answers—i.e., the individual 12 Likert-item narrative engagement scores answered for each condition—with Dunn-Bonferroni post-hoc pairwise comparisons for significant results.

[14]While there are examples both ways, we align ourselves with authors who have treated narrative engagement data as non-parametric, e.g., [37, 38, 91].

CONDITION	NARRATIVE ENGAGEMENT				
	NARRATIVE UNDERSTANDING	ATTENTIONAL FOCUS	NARRATIVE PRESENCE	EMOTIONAL ENGAGEMENT	**TOTAL SCORE**
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 N	17	15	13	14	**60**

Table 5. Median raw scores for each of the four sub-scales and median total Narrative Engagement score. See Figure 7 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.

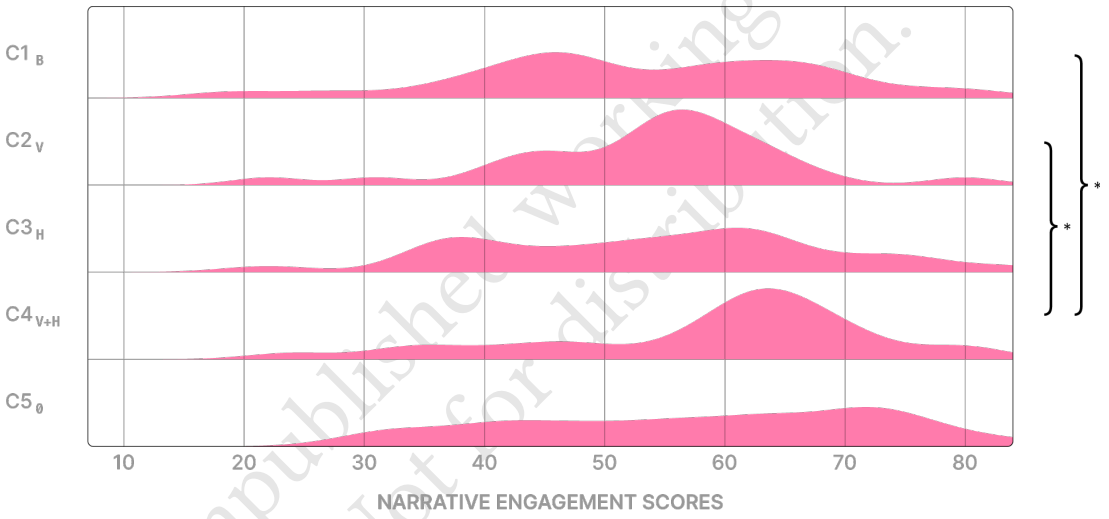


Fig. 7. Ridge plot of Narrative Engagement scores across conditions, ranging from 7 to 84. Each ridge represents a condition, with its height indicating the density of scores. Significant pairwise comparisons ($p < 0.05$) between C4 V+H and C2 V, and C4 V+H and C1 B, are highlighted by the curly brackets.

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.

5.2.2 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.

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 $C4_{V+H}$, 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 visuals "would work great in a scene where there were multiple people talking with differing emotions and tones. The colored words (C2–5) and emboldened fonts ($C2_V$ and $C4_{V+H}$) 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 $C2_V$ 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, $C2_V$ allowed them to see the character's emotions, which helped them "get more connected to the speaker's voice." P12 thought $C2_V$ 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 ($C1_B$), some participants shared that experiencing the other modalities made conventional captions feel lacking. P10 thought that going back to $C1_B$ "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 $C1_B$ 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 $C4_{V+H}$, 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 $C2_V$ 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 $C3_H$'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 $C3_H$, 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 $C4_{V+H}$ 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."

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 $C2_V$ 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 $C2_V$ 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 $C3_H$, 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 $C2_V$, 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."

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 $C5_N$ 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 $C5_N$ 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."

5.3 Discussion of Study 2

5.3.1 Using Haptic Patterns to Convey Arousal. We saw that the fourth condition ($C4_{V+H}$) stood out as the most effective. This condition combined the winning haptic pattern from Study 1 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 ($C2_V$), which was designed to mirror previously discussed affective captioning models, e.g., [25, 41, 49]. 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 RQ2, 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 ($C1_B$), i.e., conventional, non-styled captions. In other words, the $C4_{V+H}$ option was more engaging than both the conventional captions in everyday use and the recommended option from prior work on affective captions ($C2_V$).

5.3.2 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,[15] 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. [85], who previously combined haptics and captions to aid with speaker identification. While this diversity in users' experiences could simply be

[15]Spatial 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.

a byproduct of the inherent diversity within the DHH population in general [77], the tension between enhancement and distraction also aligns with broader challenges in designing multimodal captioning systems [9], particularly when considering the cumulative effects of such features over longer durations.

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 [47], 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 [24], or with how certain haptic stimuli may become more favored through repeated exposures [46]. 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 [4, 66], 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 [8]. We echo Kushalnagar et al. [51], who found that visual-tactile captions for non-speech information outperformed tactile-only ones. This effect may explain the non-significant advantage[16] of $C4_{V+H}$ over $C3_{H}$, where haptics and visuals outperformed haptics alone for conveying arousal. The higher performance of $C4_{V+H}$ over $C3_{H}$ 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 $C5_{N}$—which had *no* depiction of arousal—over both $C2_{V}$ and $C3_{H}$ 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 [28]. 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 $C4_{V+H}$ 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

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

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 [85]. 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?

5.3.3 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., [24, 41]. 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., [24, 25, 49]. Still, there were cases where the font-color and weight modulations did not work well.

The color palette recommended by Hassan et al. [41] 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 [10, 43], some researchers have advocated that design solutions are made to be purposefully ambiguous and open to interpretation [34]. 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.

Prior work has suggested the use of changes to font-weight to depict arousal, but there are no specific guidelines for how these should be implemented [24]. While minor changes in weight may not significantly affect legibility [64], 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. [58], 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.

6 Limitations

de Lacerda Pataca et al. [24] suggested modulating either font-size or font-weight to convey arousal. 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 [29, 85]. 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.

7 Conclusion

In this paper, 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 1, 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 2, 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 $C4_{V+H}$, 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 ($C1_B$) 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 ($C2_V$).

Haptics have shown promise as an addition to affective captions, with noticeable improvements in narrative engagement with audio-visual content among DHH individuals. Furthermore, participants consistently favored the low-frequency SSP haptic pattern. Our results suggest a combination of visual cues and haptics as a promising option for affective captions.

Acknowledgments

This material is based upon work supported by the Fulbright Commission (Fulbright-CAPES Scholarship, ME / CAPES Nº8 / 2020), the National Science Foundation under Grants Nº 1954284, 2125362, 2212303, and 2235405, and Department of Health and Human Services under Grant Nº 90DPCP0002-0100.

We acknowledge the use of large language model software as an editorial tool to enhance the clarity and style of this manuscript.

We thank Yiwen Wang for sharing the STL files for 3d-printing the casing for the Vibro-Transducer and Emily Kuang for reviewing an earlier version of the paper.

References

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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.

[7] 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

[8] 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.

[9] 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.hap20.153.55odbcba

[10] 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 (Aarhus05)*. ACM, New York, NY, USA, 59–68. https://doi.org/10.1145/1094562.1094570

[11] 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

[12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[13] 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

[14] Rick Busselle and Helena Bilandzic. 2009. Measuring Narrative Engagement. *Media Psychology* 12, 4 (Nov. 2009), 321–347. https://doi.org/10.1080/15213260903287259

[15] João Couceiro e Castro, Pedro Martins, Ana Boavida, and Penousal Machado. 2019. Máquina de Ouver-From Sound to Type: Finding the Visual Representation of Speech by Mapping Sound Features to Typographic Variables. In *Proceedings of the 9th International Conference on Digital and Interactive Arts*. Association for Computing Machinery, Braga, Portugal, 1–8.

[16] 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.

[17] Jonathan Cohen. 2018. *Defining Identification: A Theoretical Look at the Identification of Audiences With Media Characters*. Routledge, London, UK, 253–272. https://doi.org/10.4324/9781315164441-14

[18] 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* 16, 4 (2023), 1–6. https://doi.org/10.1109/TOH.2023.3275185

[19] Peter Constable, Saisang Cai, Ken Turetzky, and Mike Jacobs. 2018. OpenType specification version 1.8. https://learn.microsoft.com/en-us/typography/opentype/otspec180/otvaroverview. Accessed on March, 2024.

[20] Bruyne L. De, De Clercq Orphée, and Hoste Véronique. 2021. Annotating affective dimensions in user-generated content. *Language Resources and Evaluation* 55, 4 (12 2021), 1017–1045. https://www.proquest.com/scholarly-journals/annotating-affective-dimensions-user-generated/docview/2580827900/se-2

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

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

[23] 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

[24] 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

[25] 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

[26] 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

[27] Arpad E Elo. 1978. *The rating of chessplayers, past and present.* Arco Pub., New York.

[28] 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

[29] 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 Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300276

[30] Harvey Fletcher and Wilden A Munson. 1933. Loudness, its definition, measurement and calculation. *Bell System Technical Journal* 12, 4 (1933), 377–430.

[31] Mark D. Fletcher. 2021. Can Haptic Stimulation Enhance Music Perception in Hearing-Impaired Listeners? *Frontiers in Neuroscience* 15 (2021), 1–16. https://doi.org/10.3389/fnins.2021.723877

[32] 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.

[33] 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), 1–25. https://doi.org/10.3390/s23062968

[34] 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

[35] Daniel T Gilbert. 1991. How mental systems believe. *American psychologist* 46, 2 (1991), 107.

[36] 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

[37] 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* I41 (Jan. 2024), 230–254. https://doi.org/10.26034/cm.jostrans.2024.4724

[38] 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

[39] Kaixin Han, Weitao You, Shuhui Shi, and Lingyun Sun. 2024. Hearing with the eyes: modulating lyrics typography for music visualization. *The Visual Computer* 40, 11 (2024), 8345–8361. https://doi.org/10.1007/s00371-023-03239-5

[40] 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

[41] 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

[42] 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/

[43] Kristina Höök, Anna Ståhl, Petra Sundström, and Jarmo Laaksolaahti. 2008. Interactional Empowerment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 647–656. https://doi.org/10.1145/1357054.1357157

[44] Jiaxiong Hu, Qianyao 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

[45] 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

[46] 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

[47] 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

[48] 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

[49] 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

[50] 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

[51] 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

[52] 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

[53] 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, Cham, Switzerland, 139–185. https://doi.org/10.1007/978-3-319-10190-3_8

[54] DANIEL G. LEE, DEBORAH I. FELS, and JOHN PATRICK UDO. 2007. Emotive captioning. *Comput. Entertain.* 5, 2, Article 11 (apr 2007), 15 pages. https://doi.org/10.1145/1279540.1279551

[55] Jérôme Louradour. 2023. whisper-timestamped. https://github.com/linto-ai/whisper-timestamped.

[56] Karon E MacLean. 2008. Haptic interaction design for everyday interfaces. *Reviews of Human Factors and Ergonomics* 4, 1 (2008), 149–194.

[57] Fiona Macpherson. 2018. *Sensory Substitution and Augmentation: An Introduction.* British Academy, London, UK, 1–42. https://doi.org/10.5871/bacad/9780197266441.003.0001

[58] 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* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 380, 8 pages. https://doi.org/10.1145/3544549.3573872

[59] Sebastian Merchel and M Ercan Altinsoy. 2018. *Auditory-tactile experience of music*. Springer International Publishing, Cham, Switzerland, 123–148.

[60] 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

[61] Michael Mulshine, Ge Wang, Chris Chafe, Jack Atherton, terry feng, and Celeste Betancur. 2023. WebChucK: Computer Music Programming on the Web. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Miguel Ortiz and Adnan Marquez-Borbon (Eds.). Mexico City, Mexico, Article 28, 6 pages. http://nime.org/proceedings/2023/nime2023_28.pdf

[62] Stephen Nixon, Lisa Huang, Katja Schimmel, Rafał Buchner, and Cris R Hernández. 2023. Recursive Sans & Mono. http://www.recursive.design/

[63] Desmond C. Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. 2021. Modeling Emotion in Complex Stories: The Stanford Emotional Narratives Dataset. *IEEE Transactions on Affective Computing* 12, 3 (2021), 579–594. https://doi.org/10.1109/TAFFC.2019.2955949

[64] 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

[65] 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* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763

[66] 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

[67] 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

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

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

[70] Marie-Laure Ryan. 2007. Toward a definition of narrative. *The Cambridge companion to narrative* 22 (2007), 22–35.

[71] Tim Schlippe, Shaimaa Alessai, Ghanimeh El-Taweel, Matthias Wölfel, and Wajdi Zaghouani. 2020. Visualizing Voice Characteristics with Type Design in Closed Captions for Arabic. In *2020 International Conference on Cyberworlds (CW)*. IEEE, IEEE, Caen, France, 196–203.

[72] 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, Reston, VA, USA, 1–10. https://doi.org/10.2514/6.2015-2905

[73] Mark Seidenberg. 2017. *Language at the Speed of Sight: How we Read, Why so Many CanÕt, and what can be done about it.* Basic Books, New York, NY, USA.

[74] Hasti Seifi and Karon E. MacLean. 2013. A first look at individuals' affective ratings of vibrations. In *2013 World Haptics Conference (WHC)*. IEEE, Piscataway, NJ, 605–610. https://doi.org/10.1109/WHC.2013.6548477

[75] 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

[76] 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* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 705–714. https://doi.org/10.1145/1978942.1979044

[77] 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
[78] Juho Snellman. 2015. Win probability? https://github.com/sublee/trueskill/issues/1#issuecomment-149762508
[79] David Ternes and Karon E. MacLean. 2008. *Designing Large Sets of Haptic Icons with Rhythm.* Springer Berlin Heidelberg, Berlin, Germany, 199–208. https://doi.org/10.1007/978-3-540-69057-3_24
[80] Walda Verbaenen. 2019. *Phonotype. The visual identity of a language according to its phonology.* Master's thesis. PXL-MAD.
[81] Ronald T. Verrillo. 1992. Vibration Sensation in Humans. *Music Perception* 9, 3 (04 1992), 281–302. https://doi.org/10.2307/40285553
[82] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2023. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 1–13. https://doi.org/10.1109/TPAMI.2023.3263585
[83] Shaun Wallace, Rick Treitman, Jeff Huang, Ben D. Sawyer, and Zoya Bylinskii. 2020. Accelerating Adult Readers with Typeface: A Study of Individual Preferences and Effectiveness. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382985
[84] 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
[85] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 781, 14 pages. https://doi.org/10.1145/3544548.3581076
[86] 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
[87] 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
[88] 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
[89] 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
[90] Matthias Wölfel, Tim Schlippe, and Angelo Stitz. 2015. Voice driven type design. In *2015 international conference on speech technology and human-computer dialogue (SpeD)*. IEEE, IEEE, Bucharest, Romania, 1–9.
[91] 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

A Appendix: 12-Item Narrative Engagement Scale

The questions below, adapted from Busselle and Bilandzic [14], 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.