Fuzzy Feelings: Arousal's Interpretive Noise and the Case for Acoustic-Based Haptics

Caluã de Lacerda Pataca*
Department of Computer Science
Birmingham City University
Birmingham, United Kingdom
Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York, USA
calua.pataca@bcu.ac.uk

Stephanie Patterson*
School of Information
Rochester Institute of Technology
Rochester, New York, USA
slp9101@rit.edu

Roshan L Peiris
School of Information
Rochester Institute of Technology
Rochester, New York, USA
roshan.peiris@rit.edu

Matt Huenerfauth
Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York, USA
matt.huenerfauth@rit.edu

Figure 1: Haptic feedback—here implemented as speech-driven vibrotactile signals—can enrich captions for Deaf and Hard-of-Hearing viewers, but research has so far focused on vaguely defined speech arousal cues. We introduce an acoustic-based approach that maps speech features into distinct emotional vibrations.

Abstract

Captions rarely convey emotional nuances in speech, leaving Deaf and Hard-of-Hearing (DHH) viewers without access to tonal and affective information. We present a two-part mixed-methods study on how haptic feedback can communicate vocal emotion without adding visual load. In Part 1, we replicated an arousal-driven captioning approach using speech-emotion-recognition to modulate typographic weight and vibration intensity. Participants showed divergent mental models and often mapped "more vibration" to loudness rather than emotional arousal, underscoring the construct's conceptual fuzziness. In Part 2, we evaluated five acoustic-to-haptic mappings that bypass affective inference and translate pitch, rhythm, and waveform cues into vibration patterns. No single pattern dominated, but participants associated options such as PULSE or SAWTOOTH with high-arousal emotions, and PITCH-NORMALIZED signals with calmer states. We derive design guidelines emphasizing contrastive, acoustically grounded mappings and user control for integrating emotional haptics into short-form, captioned media.

CCS Concepts

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

Keywords

Accessibility, Emotion / Affective Computing, Individuals with Disabilities & Assistive Technologies, Empirical study that tells us about how people use a system

*These authors contributed equally to this research.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3793421

ACM Reference Format:
Caluã de Lacerda Pataca, Stephanie Patterson, Roshan L Peiris, and Matt Huenerfauth. 2026. Fuzzy Feelings: Arousal's Interpretive Noise and the Case for Acoustic-Based Haptics. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3772318.3793421

1 Introduction

Captions are central to speech accessibility, translating voiced language into written text for people with limited access to audio, such as Deaf and Hard-of-Hearing[1] (DHH) individuals. Yet, they remain an imperfect translation: despite well-established benefits [21], issues persist with speed, accuracy, and—central here—their lack of expressive and affective information [10].

Much of speech lies beyond literal words. A lack of tone, rhythm, and emotional nuance flattens expressive meaning—a known pain point for DHH viewers [10, 33]. To address this, researchers have explored richer caption forms, often via visual typographic modulations echoing prosodic[2] [5, 9, 47, 59] or affective [12, 23, 26, 48] aspects of speech. DHH viewers generally find emotional cues helpful [10], though these visual modulations conveyed valence better than arousal [12]. To work around arousal's weaker visual encoding, recent work paired captions with *haptic cues* that indicated excitement or calm; it was found that such combinations increased engagement relative to conventional captions [11].

Despite these benefits, a key question remains: what do DHH viewers actually understand arousal to be? Unlike valence, arousal lacks a stable technical definition, at times described as excitement, intensity, activation, or salience [50]. Studies have shown this ambiguity in both captioning and haptics: DHH viewers often conflate arousal with loudness [8, 11]. If arousal-driven haptics aim to support comprehension, we must assess whether the construct is meaningful for DHH viewers; if not, alternative acoustic-haptic mappings are needed.

To explore this issue, we conducted a two-part study (Figure 2) examining how DHH viewers interpret haptic cues that are either arousal-driven or acoustically grounded. Part 1 replicates and extends prior arousal-based caption-haptic work [11] to probe how viewers conceptualize these cues. Participants showed divergent, often conflicting mental models—echoing theoretical tensions in psychology [50]—raising questions about how consistently meaningful arousal is for many DHH viewers.

Given these divergent interpretations, Part 2 examines an alternative haptic mapping approach that avoids arousal inference entirely. In this part, we tested multi-parameter, acoustically grounded haptic patterns representing four discrete emotions, allowing us to assess both mapping clarity and which interpretive strategies DHH viewers use when judging affective haptics.

Across both parts, qualitative analysis revealed shared interpretive tendencies—attention to contrast, rhythm, and material[3] "feel"—shaping how participants assigned affective meaning. These themes support our broader claim: mappings tied directly to acoustic features offer clear and interpretable emotional cues that allow designers to sidestep issues related to arousal-based abstractions.

Regardless of arousal's usefulness as a construct, implementing it in real-time captioning also carries technical challenges. Speech-emotion-recognition (SER) requires computationally intensive models; although de Lacerda Pataca et al. [11] found arousal-driven haptics immersive, they used pre-processed stimuli. In real-time settings, any mismatch between visual and haptic cues may reduce or eliminate benefits, since cross-modal consonance predicts immersion [58]. Further, SER models remain vulnerable to gender, speaker, linguistic, and domain biases [54].

As an alternative approach, prior work has shown that haptic signals derived directly from a speaker's voice can convey affective qualities without explicit emotion inference [56]. Building on this insight, Part 2 explores acoustic-driven haptic mappings that bypass intermediate affective modeling altogether. We developed five acoustic-to-haptic mappings and selected four canonical discrete emotions—one per circumplex quadrant—to examine how DHH viewers interpret haptic patterns grounded in specific speech features. This design allows us to characterize the interpretive strategies participants use when engaging with acoustically grounded affective haptics, without presupposing their relationship to arousal-based cues.

Our contributions are as follows:

(1) Empirical evidence of divergent mental models of haptics and speaker arousal, highlighting ambiguity in how the construct is interpreted by DHH viewers;
(2) Novel acoustic-to-haptic mappings that translate emotional cues in speech into distinct haptic patterns for four discrete emotions;
(3) Actionable design guidelines for employing haptic patterns expressing specific emotions (*e.g.,* happiness, anger) in short-form, captioned videos.

2 Background

Here, we review limitations of current captioning and challenges in representing non-linguistic speech features. We focus on haptic representations of emotional information, examining prior work encoding affect through haptic stimulation, much of which centers on arousal as a key parameter. We end with recent critiques showing that the affective construct is difficult to operationalize.

2.1 Traditional captions as an imperfect but essential accessibility tool

For over 50 million U.S. residents with hearing loss [24], captions are essential for accessing auditory information. Since 2006, U.S. law has required captions for all new video programming [39], yet issues persist. Standards such as WCAG mandate access to non-speech information (NSI) like sound effects and music [6, 53]. Extra-speech information (ESI), a subset of NSI, covers non-linguistic aspects of speech—identity, tone, vocal affect. For instance, "[Upbeat pop music playing]" conveys NSI, while "Alex: [Cheerfully] Hello!" conveys ESI (speaker identity and emotional tone).

[1] We acknowledge the diverse identities represented in the DHH acronym, including deaf, Deaf, HOH, Black-deaf, etc.
[2] Acoustic features such as loudness, pitch, and rhythm.
[3] "Material properties" refer to signal-level features (rhythm, intensity, spectral brightness) and how they feel on the body.

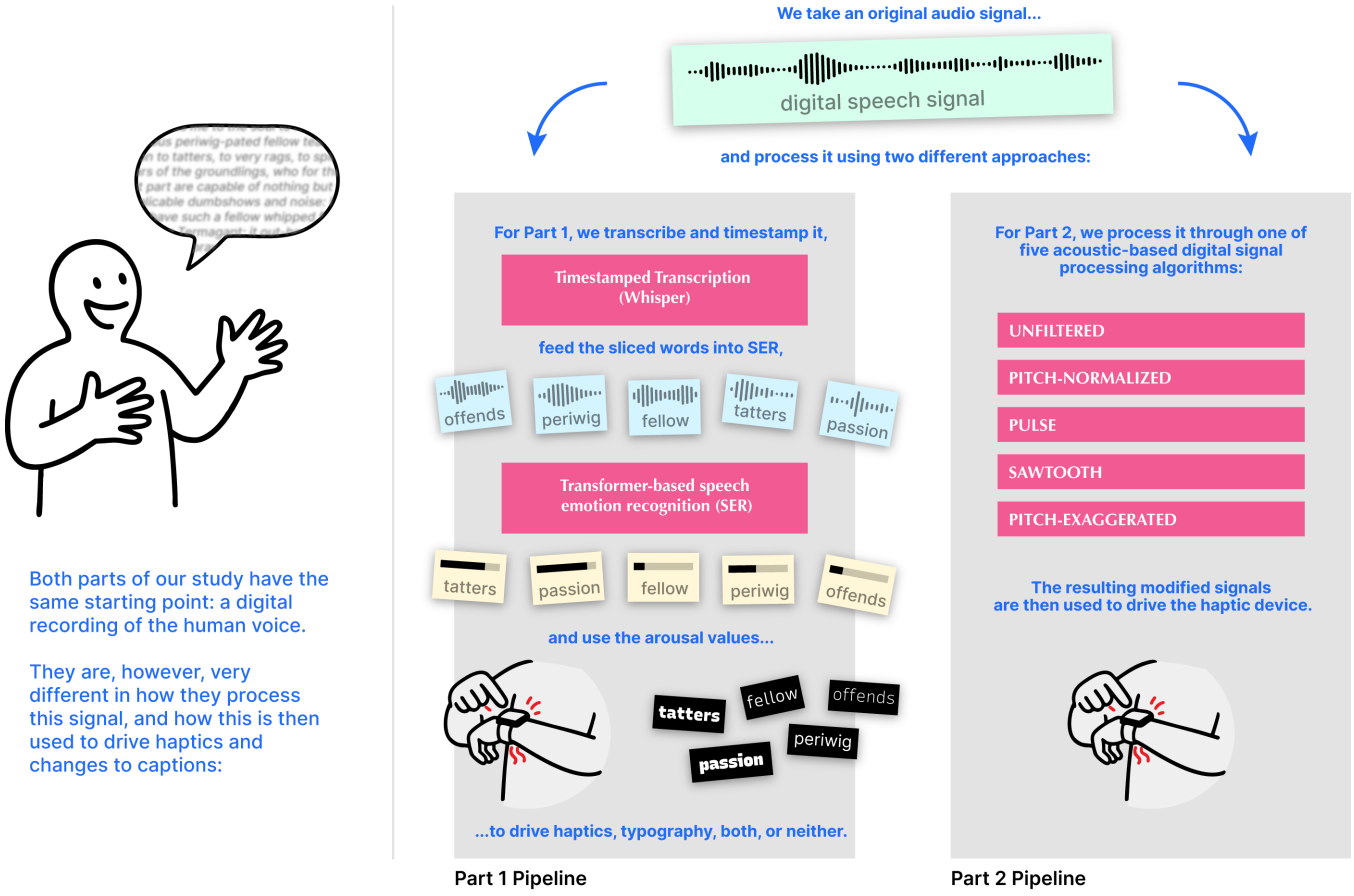


Figure 2: Overview of the two study pipelines. Both parts start from the same digital speech signal but diverge in how the signal is processed and mapped to haptics and captions. Part 1 uses timestamped transcription and transformer-based speech-emotion-recognition to infer arousal values that drive visual and/or haptic modulations. Part 2 bypasses affect inference and instead applies one of five acoustic-based signal processing algorithms to generate haptic patterns directly from the speech signal.

Yet many captions omit NSI and ESI entirely. Surveys with DHH users report minimal or low-quality depictions of NSI, causing confusion and frustration [31]. Auto-generated captions also neglect these cues [32]. Captionists and viewers alike describe NSI/ESI as underdeveloped, with limited conventions and inconsistent judgments about what to include. McDonnell and Findlater [35] further argue that captioning research often overlooks relational and contextual aspects of access, reinforcing gaps in how these dimensions are handled.

2.2 Expressive captioning and other representations of ESI

To bridge expressive gaps between captions and speech, researchers have explored techniques that visually encode prosodic and affective cues by modulating typography. de Lacerda Pataca et al. [10] identified which speech dimensions users wanted represented and found strong preference for affective information. A follow-up study [12] compared modulation styles, revealing more consistent preferences for valence-encoding than arousal-encoding cues. While such augmentations can enrich captions, readability and comprehension remain priorities for DHH users. This work suggests value in going beyond traditional captions, but not at the cost of visual overload [28], motivating exploration of complementary modalities—especially haptics—to convey affect without adding visual load.

2.3 Prior work on adding haptic feedback to captioning

To address the uneven performance of visual styles—effective for valence but less so for arousal—researchers have explored haptics as a complementary channel. de Lacerda Pataca et al. [11] extended [12] by adding a wrist-worn device encoding arousal via vibration patterns; participants reported greater engagement when haptic and visual cues were combined. Gage et al. [18] used haptics through a PS5 controller to represent NSI in action films and found improved enjoyment, clarity, and engagement. Related media-enhancement work shows haptics can shape affective interpretation: Mazzoni and Bryan-Kinns [34] found that higher-frequency, higher-intensity patterns increased self-reported arousal, and their synthetic vibration

sequences were often interpreted as echoing natural sound events during film viewing.

Although many systems map arousal to haptic intensity, evidence shows participants also interpret haptics in terms of valence and other affective qualities [11], challenging a strict visual–valence / haptic–arousal divide. Wang et al. [56] found that directly translating speech into haptics helped DHH participants distinguish voices (their primary goal) but also perceive affective qualities. Other work shows DHH individuals can distinguish sound and pitch from audio transformed into haptic patterns [17, 41]. This suggests translating emotion-correlated acoustic features [22] into haptics is promising, but designers must consider not only what signals are delivered but how they are interpreted.

2.4 Interpretation of haptic signals

Communicating information through haptics is not straightforward. Zhou et al. [60] identify a tension between devices designed for clarity and those that invite ambiguity. Rather than a flaw, some argue ambiguity supports sense-making [20]. This openness aligns with ESI's interpretive nature [40] but contrasts with affective computing traditions that treat emotion as objectively conveyable [42]. We follow the view that meaning is enacted in context [4]: interpretation depends on relational, situational, and cultural embeddings. The same cue can read differently—*e.g.,* a strong vibration might suggest anger or simply vocal effort in a noisy club—and open-ended designs can lead users to invent their own affective encodings [51].

This open-endedness reflects embodied associations: touch mediates our relation to the world, and prior tactile experiences shape perception. Haptics can also bias perception across modalities. For instance, Gatti et al. [19] found haptic feedback altered arousal ratings of images; Salminen et al. [45] showed amplitude-based haptic echoes heightened perceived arousal; and Ackerman et al. [1] demonstrated that incidental sensations (*e.g.,* weight or hardness) biased social judgments. Similarly, Turchet et al. [52] showed vibrotactile music enhanced immersion, arousal, and emotional interpretation for cochlear implant users. Such effects highlight that meaning emerges from how a signal is situated and experienced, complicating attempts to tie haptics to specific emotional constructs—especially arousal, whose boundaries remain contested.

Such effects support the view that sensorimotor experience scaffolds conceptual knowledge, with later tactile encounters drawing on these embodied metaphors [1]. These observations underscore that haptic meaning does not reside in the signal alone, but emerges from how it is situated and experienced. These complexities complicate attempts to tie haptic patterns to specific emotional constructs—especially arousal, whose conceptual boundaries have long been contested.

2.5 Issues with arousal as a construct

The intertwining of sensation and meaning echoes longer debates about arousal itself, from Kant's eighteenth-century division of affects into either energizing and sedating, to the twentieth-century emergence of "emotional excitement" as a felt activation of the body [13]. The tug-of-war between physiological and phenomenological views helps explain why arousal remains ill-defined [3] and raises a practical challenge: how should we treat the construct when designing haptic captions?

Researchers urge caution. Smith et al. [50] argue that arousal has become a "fuzzy" umbrella spanning related but distinct phenomena, with no unequivocal physiological correlates and inconsistent perceptual findings. Because the term is entrenched, they recommend replacing vague usage with more explicit constructs. Sander [46] further proposes abandoning *unspecified* usage and reframing many cases as *affective relevance*—context-dependent salience that guides attention and action.[4]

2.6 Research Questions

Together, these threads reveal gaps in how captions encode affect—particularly arousal—through haptics and typographic means. First, although (1) haptics have been shown to influence arousal judgments in audiovisual stimuli and (2) speech-synchronous haptic feedback can increase engagement, little is known about how effectively viewers, especially DHH individuals, decode such signals. As such, we ask:

RQ1 How well do modulations of haptics and typography map to how DHH participants perceive and conceptualize a speaker's arousal levels?

Answering RQ1 addresses the *clarity* of a combined haptic / typographic design. It does not address how haptic material qualities influence perceived affect, the focus of RQ2:

RQ2 What are the material qualities in a haptic signal that best correlate to the perceptual/phenomenological qualities in speech with distinct emotional tones?

3 Part 1: Investigation of DHH individuals' perceptions of arousal as an affective construct

We ran an in-person study with DHH individuals who watched captioned videos while wearing a wrist-mounted haptic device (Figures 4a–4b). This particular form factor follows from Findlater et al. [15]'s findings that smart-watches were the most preferred locus for haptic feedback in sound-awareness systems, with participants commonly selecting wrist-based haptics in their ideal configurations. The study involved two sequential tasks (Figure 3). Although we describe procedures separately, our qualitative analysis integrates open-ended responses from both parts: participants used similar interpretive strategies when reasoning about arousal-driven cues in Part 1 and about acoustic–haptic patterns in Part 2, motivating a unified thematic analysis (Section 3.2).

3.1 Methods for Part 1

Methodologically, Part 1 closely follows de Lacerda Pataca et al. [11] in using an SER-derived arousal signal to drive caption and haptic modulations, but differs in using shorter, more diverse videos and avoiding immersion measures. This shift allowed us to focus on participants' interpretive strategies for speakers across a broader set of settings.

[4]Here, we use "arousal" pragmatically as the output of a speech-emotion-recognition model that aggregates acoustic cues into a [0,1] estimate of how "energetic" a segment sounds.

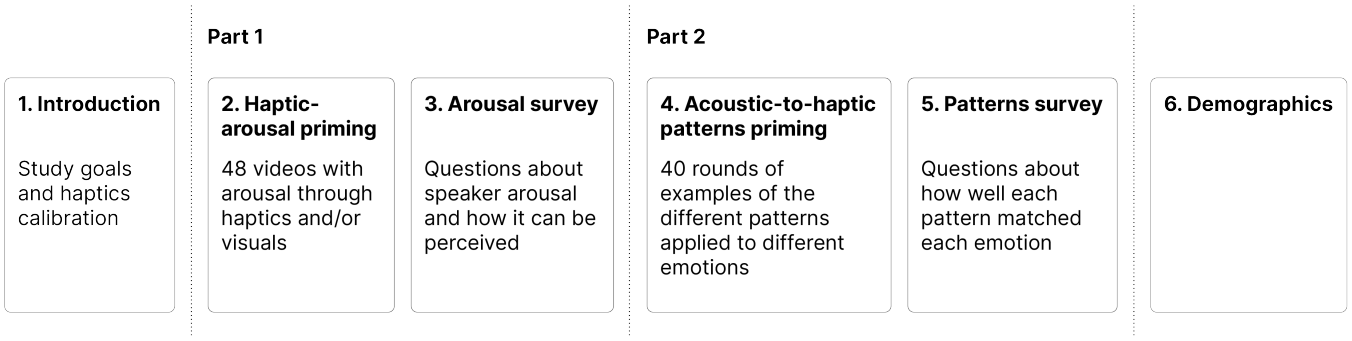


Figure 3: Overview of all experimental steps. Part 1 includes haptic–arousal priming and an arousal-rating survey using 48 captioned video clips. Part 2 includes priming and evaluation of five acoustic-to-haptic mappings across 40 rounds. Demographic questions were collected at the end.

Table 1: Participant demographics.

PID	Gender	Age	Status	CI	Primary Comm.
P1	F	58	HoH	Yes	Lipreading
P2	F	38	d/Deaf	Yes	Speech + ASL
P3	F	43	HoH	Yes	English + ASL
P4	F	49	d/Deaf	No	ASL
P5	F	54	d/Deaf	Yes	Speech
P6	F	20	d/Deaf	No	Spoken English
P7	F	19	d/Deaf	Yes	ASL
P8	M	27	d/Deaf	Yes	ASL
P9	F	47	HoH	Yes	Spoken English + ASL
P10	M	24	d/Deaf	Yes	Sim-com, ASL
P11	F	55	Deaf (1 ear)	Yes	Verbal
P12	M	28	d/Deaf	Yes	English
P13	F	65	Deaf / LV	No	ASL
P14	F	72	d/Deaf	Yes	Oral

3.1.1 Study design. An ASL-proficient author facilitated the study, which received Institutional Review Board approval. Participants were recruited through local Facebook groups and snowball sampling, with sessions held in a research lab or library room. Eligibility required self-identifying as d/Deaf or Hard-of-Hearing. Fourteen participants enrolled (11 female, 3 male; 9 d/Deaf, 3 Hard-of-Hearing, 2 other: deaf in one ear; deaf & low-vision; see Table 1), mean age 43 (σ=17), and received $40 compensation.

Participants first calibrated the amplifier so the lowest amplitude produced a *just-noticeable* vibration and the highest a *comfortable maximum*. After the facilitator explained Part 1, they completed a priming task with 48 clips showing arousal via caption modulations, haptics, both, or neither (details in Section 3.1.2). Stimuli order and condition assignment were counterbalanced. Across all trials, participants watched captioned clips while arousal cues were factorially varied. They focused on arousal intensity and how each setting shaped interpretation; these ratings served primarily as elicitation and priming instruments rather than as a reliable quantitative measure of perceived arousal. Participants then answered open-ended questions about how they infer arousal in conversation, which cues signal excitement or calmness, and their confidence in these judgments. After the second priming task (Section 4), they described what makes a haptic pattern match an emotion and what signals guided their interpretations.

3.1.2 Arousal inference. For the first priming task, we adapted the conditions from de Lacerda Pataca et al.'s study [11]. Because our focus was on arousal rather than valence, we excluded the latter: the documented U-shaped relationship between valence and arousal [27] would make the former a confounding variable in our analysis. Videos were shown under four conditions: conventional captions, visual arousal cues, haptic cues, and combined visual-haptic cues.

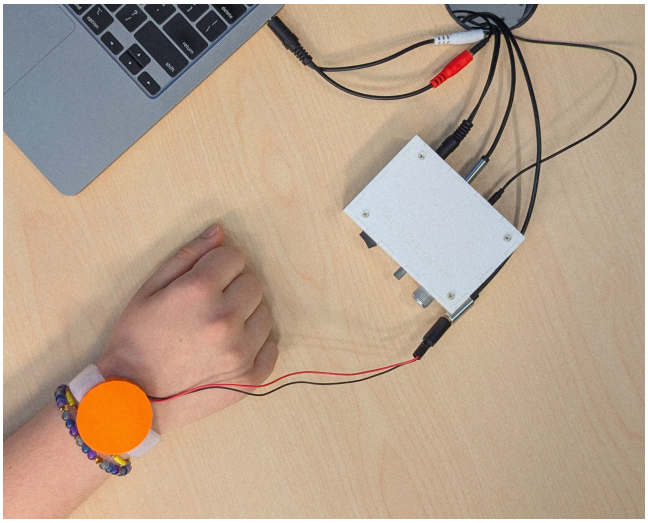
We collected short (2–12 s) video clips from mainstream streaming and video-sharing platforms, including talks, interviews, films, online videos, and recorded video calls. Clips were chosen to ensure diversity in speech style (natural vs. acted), genre, and speaker age and gender. We prioritized single-speaker scenes with clean audio that avoided competing sound sources. Clips were transcribed with Whisper [43] and its timestamping extension [30]. Each word token was fed into a transformer model [54] to infer arousal, encoded as WebVTT classes (*e.g.,* `v.arou_0p565` for a word with 0.565 inferred arousal) [7], which our typography and haptic pipelines read.

To select the final excerpts, we sampled short segments spanning a broad arousal range, seeking representative high- and low-arousal speech. We favored locally stable passages whose mean arousal values evenly covered the [0,1] range, yielding comparable clips representing calm, moderate, and highly energetic speech.

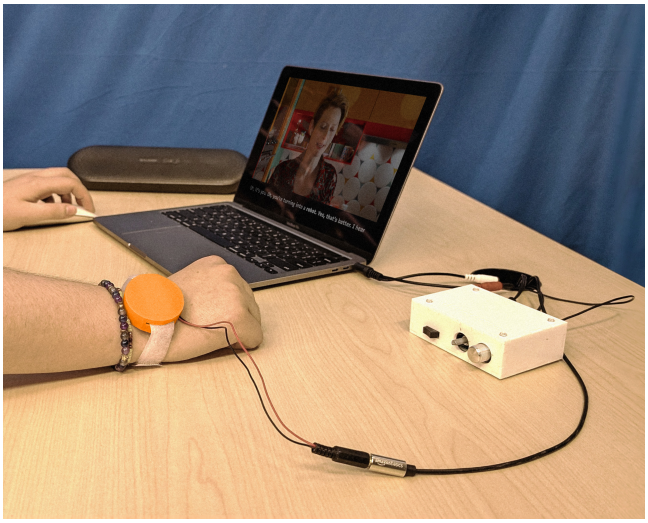
3.1.3 Depicting arousal through haptic feedback. Following prior work [11, 56], we delivered haptic feedback using Acouve's Vp2 Vibro-Transducer[5], powered by a Techtile Toolkit amplifier [36] and enclosed in a 3D-printed wrist case (Figure 4a). The device functions as a loudspeaker driver transmitting vibrations through direct contact and is driven by an audio-interface → amplifier → actuator chain.

We generated the signals by reading per-word arousal values from the WebVTT captions and modulating the intensity of a 75 Hz

[5]We did not obtain direct mechanical measurements of vibrotactile output (*e.g.,* RMS acceleration, displacement, or frequency response), as we lacked access to specialized vibrotactile instrumentation. According to the manufacturer's specifications, the Acouve Vp2 vibro-transducer supports tactile frequencies starting at approximately 16 Hz and reproduction from 20 to 15,000 Hz. The device has a diameter of 43 mm, thickness of 15 mm, and weighs 49 g (http://www.acouve.com/product/prod_vp2.asp).



(a) Close-up of the Techtile Toolkit vibrotactile driver used in both studies. The photograph shows the amplifier box, audio input connection from the laptop, and the wired output to the transducer worn by the participant.



(b) Participant wearing the wrist-mounted vibrotactile transducer while watching captioned video clips on a laptop. The Techtile Toolkit driver (right) delivers synchronized vibration patterns to the wrist as the clips play.

Figure 4: Experimental setup used in both studies, showing the vibrotactile driver and the wrist-mounted transducer during video viewing.

sine pulse lasting each word's duration. This replicated the top-performing pattern in de Lacerda Pataca et al., with a bell-shaped pulse covering two-thirds of each word followed by a brief pause to create discrete haptic units. Signals were synthesized in ChucK [55] using WebChucK [37] and mixed so original audio played through speakers while the haptic signal was routed to the device (Appendix A, Algorithm 1).

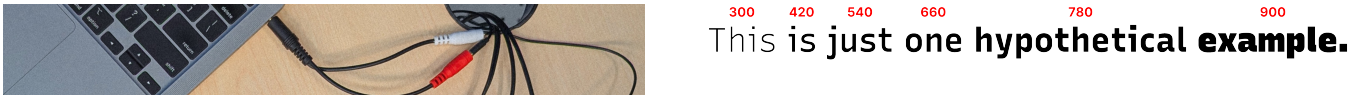


Figure 5: Example of font-weight modulation from the lowest (300) to the highest (900) possible values in the Recursive typeface. Per-word font-weight values appear in red.

3.1.4 Depicting arousal through typographic modulations. Two conditions conveyed per-word arousal through visual changes. We modulated each word's font weight, a well-established approach for expressing affective and acoustic cues in captions [8, 9, 12, 59]. Words appeared thinner or thicker depending on arousal values. Following de Lacerda Pataca et al. [11], we used the Recursive typeface [38], whose variable weight axis spans 300/thin–900/thick (Figure 5).

3.2 Results and Discussion of Part 1

To understand how DHH viewers make sense of affective haptics, the thematic analysis synthesizes qualitative responses from both parts of the study: (1) reflections on arousal-based haptics and expressive captions in Part 1, and (2) participants' evaluations of the five haptic patterns in Part 2 (Section 4).

Two authors independently coded the data and iteratively developed themes based on codebook similarity. The analysis revealed differences in how participants understood arousal and how they believed it could be represented through haptics. These differences shaped how they mapped haptic signals to emotion, judged multimodal clarity, and envisioned real-world use.

3.2.1 Theme 1: Arousal as an ambiguous and inconsistently understood construct. Although arousal is a central dimension in affective science, participants did not share a common understanding of what it means for a speaker to sound "aroused." They drew on diverse and often incompatible cues to infer emotion. Some grounded their judgments in visual behaviors such as "facial expressions and body language" (P6) or noticing when "mouth movements [are] more prominent" (P2). Others relied on semantic or narrative content, attending to "the feeling behind the words" (P11) or "the topic of what they're talking about" (P4). Still others referenced broad, subjective impressions like "high energy" (P8) or even cinematic framing, such as "the way the camera focuses on them" (P4).

Many participants described arousal as "more of something"—more movement, intensity, or emphasis—but their definitions as to what counted as "more" varied widely. No single cue was reliable across users, making arousal difficult to define and even harder to map to a single perceptual channel.

Cross-cutting implication #1: Arousal is a fuzzy target. Open-ended responses showed *diverse mental models*: some participants mapped stronger / denser vibrations to "anger" or "excitement," others longer/steady vibrations to "neutral," and others used figurative descriptions like a "bouncy" signal for happiness. This heterogeneity is not surprising: conceptually, *arousal is neither unitary nor cleanly separable from valence,* with definitions that are at times circular [50].

Participants implicitly reframed arousal as *affective salience or contrast* (*i.e.,* changes over time), not a monotonic "more amplitude

→ more arousal" mapping. The latter reflects an acoustically centered view in which arousal and volume are assumed to dominate the speech signal, *e.g.*, high loudness/high pitch → high arousal [22], and is in line with previous speech-accessibility approaches [11]. By contrast, the salience perspective explains why a strong yet continuous signal read as neutral: despite high energy, low variability conveyed little contrast, aligning with accounts defining arousal as affective salience [50].

3.2.2 Theme 2: Haptic meaning-making is subjective, phenomenological, and not linearly mapped to arousal. Participants did perceive emotional information in the haptic signals, but their interpretations were not uniform. While some inferred that stronger vibrations indicated high arousal, and weaker ones calmness, others assigned meaning based on rhythm, contrast, or overall "feel." Examples included anger being described as "long intensity" (P11), or vibrations that are "dense and strong" (P3) or "harder" (P10). Happiness, by contrast, was "short with multiple durations" (P11), or "bouncy" and apparently "lighter" (P10). For P3, emotion emerged from the "urgency of the beats/taps." These abstract sequences often evoked familiar sonic events, echoing prior findings that synthetic haptics evoke rhythms and ambient sounds [34].

A recurring motif was *contrast*. Abrupt changes ("on and off" for anger, P4), lack of variation ("no change" for calm/neutral, P11), or rhythmic alternation ("bounces back and forth," P10) shaped affective judgments. Participants often described the raw, unmodulated speech signal in Part 2 as low-contrast, suggesting that haptic contrast is important for signaling emotionally rich moments, though not necessarily for representing arousal.

Cross-cutting implication #2: Emotional interpretation is individual and context-dependent. Participants stressed that the usefulness of arousal-based haptics depends on personal preference and context. Some valued the added emotional information, while others—shaped by long-term caption use—found haptics optional or unnecessary. Many said such cues "should be up to the user to turn on or off," and their suitability varied by environment: helpful in entertainment or teleconferencing but less so in clinical settings, public spaces, or situations requiring discretion. This variability reinforces that arousal is not universally meaningful and that these systems need to support customization, a common theme in accessibility and haptics research [11, 52].

3.2.3 Theme 3: Cross-modal consistency and clarity are crucial for trust and comprehension. Participants repeatedly emphasized that emotional cues—visual, semantic, and haptic—must align. Conflicts between haptics and facial expression, body language, or caption style caused confusion and reduced trust. As P5 noted, vibrations "helped when the emotions were more intense" but were otherwise confusing. P13 echoed this: "I felt conflict between the vibrations and the words. I didn't know which to trust…my gut about the meanings of the words? Or the vibrations?"

When haptics aligned with perceived affect, participants reported increased access to emotional information and reduced cognitive effort. One noted that "the vibration gave me more information about the emotions and tone" (P8); another said that without visual modulations they could focus more on facial expression and body language, suggesting that arousal-related haptics were less visually intrusive (P3).

Cross-cutting implication #3: Cross-signal consonance matters. Participants emphasized that haptic cues must *work with* visual, semantic, and contextual information. Misalignment between channels created uncertainty about which signal to trust. Users interpreted vibrations *relationally*, consistent with Wirth et al. [58], who note that immersion declines when modalities conflict. This does not preclude ML-based emotion inference (*e.g.*, [11]) but suggests such approaches may be better suited to non–real-time settings, where multimodal consistency can be ensured.

4 Part 2: Representing categorical emotions using five acoustic modulation patterns

In Part 2, we shift from SER-inferred values to five speech-derived haptic patterns that manipulate pitch, duration, and tone—paralinguistic cues that convey emotional nuance [22] and map readily onto haptic properties [2].

Participants played short acted utterances with captions while feeling the corresponding haptic pattern on the wrist. As in Part 1, the session began with a priming phase in which participants heard utterances transformed by one of the five modulation algorithms, then identified which emotion each haptic rendering best matched.

Given Part 1's evidence of limits in the interpretability of arousal as an affective dimension, in Part 2 we adopted a discrete, categorical framing that we believed would better approximate participants' intuitions about emotional qualities in speech and haptics. Thus, we did not explicitly model arousal and, consistent with its exclusion in Part 1, also did not include valence. However, the discrete categories we selected inherently encode both arousal and valence polarity, which allowed us to examine how participants mapped haptic patterns onto affective qualities without requiring continuous dimensional judgments. As such, all speech samples used the same sentence spoken in three Ekman emotions—happy, sad, angry [14]—plus calm, thus covering the four quadrants of the circumplex model through four distinct emotions [44]. Audio files were sourced from the RAVDESS Emotional Speech Audio dataset, produced by a professional actor [29].

The five haptic patterns each interpret the audio signal differently, emphasizing distinct acoustic features derived from speech: PITCH-EXAGGERATED, PULSE, SAWTOOTH, PITCH-NORMALIZED, and UNFILTERED (the original signal played as haptic vibration).

4.1 Modulation algorithms

We iteratively developed four patterns translating key acoustic indicators of speaker emotion, modulating parameters that had been kept fixed in Part 1 while ensuring that each pattern produced a distinctive haptic sensation. Instead of holding frequency, waveform, or rhythm constant (with amplitude alone reflecting arousal), we allowed these parameters to mirror acoustic changes in the speech signal. This approach directly correlates sound and haptics and is feasible for real-time use, *e.g.,* teleconferencing, whereas Wagner et al. [54]'s SER model is limited to pre-recorded content due to computational latency.

Table 2 summarizes the four modulation patterns plus UNFILTERED, which acted as a baseline and is described in Appendix B, Algorithm 3. We built an algorithm that extracts pitch, rhythm,

and amplitude from the speech signal using Python's Parselmouth library [25] (Appendix B, Algorithm 2). Because the perceptible range of haptics differs from aural speech, pitch was normalized to vibrotactile thresholds [16]. Extracted features were sent to ChucK [55], which synthesized haptic patterns approximating the pitch changes and rhythm of the original audio.

The four patterns are as follows:

- PITCH-NORMALIZED: smooths and normalizes pitch to 80–300 Hz, approximating the range of maximum vibrotactile sensitivity on the wrist [61]; (Appendix B, Algorithm 4);
- PULSE: inserts silence intervals between pitch changes; (Appendix B, Algorithm 5);
- SAWTOOTH: applies the pitch-normalization algorithm but uses a sawtooth wave instead of a sine wave; (Appendix B, Algorithm 6);
- PITCH-EXAGGERATED: raises incoming pitch values above 120 Hz, amplifying higher pitches and pitch changes; (Appendix B, Algorithm 7).

4.2 Results

4.2.1 Discrete Emotion Judgments of the Acoustic-Haptic Cues. Participants replayed each of the five haptic patterns and rated, on a 7-point Likert scale, "how well the pattern represented the emotion anger / happiness / sadness / calmness." These within-subject ratings (pattern × emotion × participant) underpin our analyses.

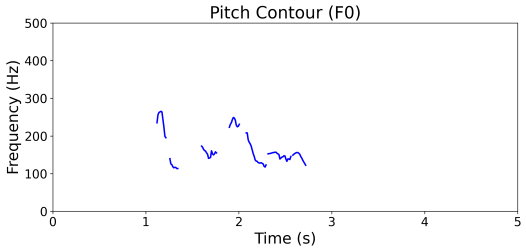
We ran Friedman tests per emotion comparing ratings across the five conditions. Omnibus effects were significant for anger, $\chi^2(4) = 23.20$, $p < .001$; calmness, $\chi^2(4) = 13.80$, $p = .008$; happiness, $\chi^2(4) = 18.39$, $p = .001$; and sadness, $\chi^2(4) = 9.77$, $p = .045$, all with $n = 14$ matched participants. Effect sizes, quantified via Kendall's W, indicated small-to-moderate concordance across conditions (anger: $W = .38$; calmness: $W = .18$; happiness: $W = .29$; sadness: $W = .15$). Median scores for the five conditions × the four emotions can be seen in Table 3.

Post-hoc Wilcoxon signed-rank tests with Holm correction revealed several significant contrasts and trend-level effects ($p < .10$). For *anger*, ratings in the PULSE condition were significantly higher than those for calmness and sadness, with trends suggesting SAWTOOTH anger ratings exceeded those for calmness and happiness. For *happiness*, PULSE was rated higher than calmness and sadness, and happiness was rated higher than sadness. *Calmness* received higher ratings than happiness under UNFILTERED, but was consistently rated lower than sadness under both PITCH-NORMALIZED and PULSE. No single condition dominated, but these contrasts illustrate how haptic patterns shaped judgments. See Table 4 for full details.

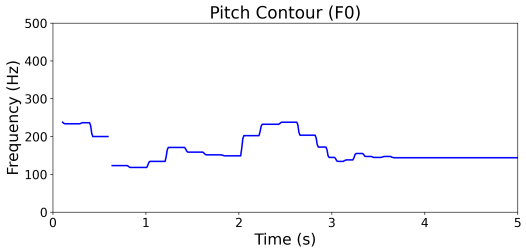
5 Discussion

5.1 Direct acoustic → haptic mappings supported affective interpretation, but were not universally decisive

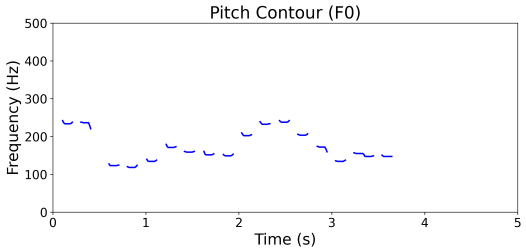
In Part 2, *no single algorithm dominated* across discrete emotions, though some associations emerged (SAWTOOTH for anger, UNFILTERED for calm, PULSE for high arousal). This supports the use of multi-parameter bundles—*i.e.,* combinations of intensity, rhythm, pitch, and waveform texture—as a viable design direction for haptic



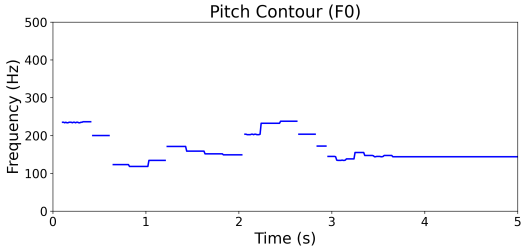
(a) Angry UNFILTERED



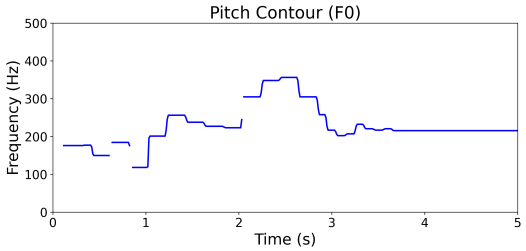
(b) Angry PITCH-NORMALIZED



(c) Angry PULSE



(d) Angry SAWTOOTH



(e) Angry PITCH-EXAGGERATED

Figure 6: Graphs generated using Python showing the pitch contour of the angrily-spoken utterance under our five different processing conditions.

Table 2: Descriptions of the five haptic conditions for the second part.

Condition	Description	Parameter modulated	Pitch contour graph
UNFILTERED	The raw speech signal	Unfiltered	Figure 6a
PITCH-NORMALIZED	Speaker pitch normalized to haptic range	Rhythm, pitch	Figure 6b
PULSE	Short tones instead of long	Length/rhythm	Figure 6c
SAWTOOTH	Sawtooth waveform instead of sine	Tone	Figure 6d
PITCH-EXAGGERATED	Speaker pitch exaggerated	Pitch	Figure 6e

Table 3: Median ratings by condition for each emotion.

Emotion	SAWTOOTH	PITCH-EXAGGERATED	UNFILTERED	PITCH-NORMALIZED	PULSE
Anger	5.5	3.0	1.0	4.5	5.0
Calmness	1.0	1.5	4.5	1.0	1.0
Happiness	3.0	3.5	1.0	3.0	4.0
Sadness	4.0	4.0	2.0	3.0	2.0

Table 4: Pairwise Wilcoxon signed-rank comparisons after Holm correction. Effects with $p < .05$ are marked and those with $p < .10$ labeled as trends. High-arousal emotions are shown in SMALL-CAPS, low-arousal ones in *italics*, and the higher emotion in each comparison in bold.

Condition	Comparison	W	p_{holm}	Higher	Note
SAWTOOTH	**ANGER** > HAPPINESS	14.5	.072	ANGER	Trend
	ANGER > *calmness*	13.0	.065	ANGER	Trend
	sadness > *calmness*	9.0	.069	*sadness*	Trend
PITCH-EXAGGERATED	Omnibus not significant; post-hoc skipped.				
UNFILTERED	ANGER < ***calmness***	7.0	.035	*calmness*	*
	HAPPINESS < ***calmness***	4.0	.024	*calmness*	*
PITCH-NORMALIZED	**ANGER** > *calmness*	5.0	.018	ANGER	*
	sadness > *calmness*	4.5	.022	*sadness*	*
PULSE	**ANGER** > *sadness*	3.0	.008	ANGER	*
	ANGER > *calmness*	1.0	.002	ANGER	*
	HAPPINESS > *sadness*	14.5	.050	HAPPINESS	*
	HAPPINESS > *calmness*	0.0	.006	HAPPINESS	*
	sadness > *calmness*	8.0	.050	*sadness*	*

affect representation. In contrast to the approach of Part 1 and prior work [11] that emphasized single-parameter, monotonic mappings, Part 2 explored how distributing information across multiple parameters shapes interpretation. Notably, waveform texture proved meaningful, contrasting with Akshita et al. [2], who viewed it as a less discernible haptic parameter. Overall, the results favor moving beyond single arousal "dials" toward a composite vocabulary that emphasizes cross-parameter contrast.

5.2 Contextual considerations

Participants cited haptics as helpful when faces are not visible (*e.g.,* narration), lip-reading is not possible, or expressions are muted. They also saw haptics as reducing load on the already overburdened visual channel. Distributing information across modalities can aid comprehension even when cues are partly redundant, aligning with Wickens' multiple resource theory, which predicts better time-sharing when tasks draw on distinct resource pools [57].

Participants also emphasized user control, noting that many would still prefer traditional captions. Those skeptical of expressive captioning agreed it fit entertainment contexts better than serious ones like work meetings or medical appointments. One participant worried that expressive captions could be visible to others, potentially marking users—an established concern in accessible technology [49]. Although smartwatches have been well received as Sound Communication Technology [15], designers must still consider how added features may draw unwanted attention or reveal unwanted information.

5.3 Design guidelines

Based on our mixed-methods results, we offer preliminary design guidelines for haptic representations of affect. These guidelines reflect patterns observed across our studies rather than a definitive comparison between mapping approaches. Content creators can use these to create rich emotional feedback for DHH users and others, since haptics may capture felt qualities of speech relevant to affect.

(1) Conceptualizing "arousal" as affective contrast or salience uncouples it from "acoustic loudness," making it more actionable for haptic design. Rhythm, waveform, and pitch can denote altered emotional states without relying on monotonic intensity-arousal mappings;
(2) A categorical model of emotions can support distinct, multi-parameter mappings, reducing reliance on strictly monotonic "more is more" assumptions. For these, we offer the following specific recommendations:
   (a) For low arousal emotions such as calmness and sadness, we recommend the PITCH-NORMALIZED haptic pattern;
   (b) For high arousal emotions such as anger or happiness, we recommend the PULSE haptic pattern;
   (c) For representing high arousal, low valence emotions such as anger, we recommend the SAWTOOTH haptic pattern;
(3) Designers should avoid haptics that conflict with visual cues (facial expressions, body language, or the captions themselves), as conflict consistently caused confusion;
(4) Additional cues like haptics are especially important when facial expressions or body language are absent. When a speaker's face or mouth is not visible, haptics can substitute for affective information typically gathered visually; when affect is visually available, haptics can still reinforce it.

6 Limitations

Our study has limitations that bound interpretation. **SAMPLE.** The sample was modest (n=14) and skewed older (μ=43, σ=17), which may limit generalizability within the DHH community and across cultural/linguistic contexts. Although we did not conduct a formal a priori power analysis we report Kendall's W effect sizes in Part 2 to contextualize the magnitude of observed differences. **STIMULI & TASK.** We used short edited clips and, in Part 2, a single acted utterance; longer, naturalistic speech may elicit different patterns. While our findings seem appropriate for short-form video, they should be tested with longer content where distraction, fatigue, and habituation to vibration may emerge. **OUTCOMES.** Our primary outcomes were judgments (Likerts, pairwise choices) and self-reports. We did not test downstream comprehension, recall, or task performance, so functional benefits remain to be established. We also did not measure whether participants would prefer to watch captioned content with these haptics enabled or disabled; a controlled preference study comparing captions with vs. without haptics remains an important next step. **COMPARABILITY ACROSS PARTS.** We did not directly compare the arousal-to-haptics mapping (Part 1) with the acoustic-to-haptics mappings (Part 2) within the same paradigm; differences across parts should not be over-interpreted as head-to-head performance. **ECOLOGICAL VALIDITY.** Sessions were in-lab. Real-world use (teleconferencing, mobile viewing, public spaces) introduces noise, divided attention, social acceptability concerns, and device heterogeneity that we did not study. These constraints suggest caution in generalizing beyond our setup and motivate future work on longer-form content, head-to-head comparisons of mapping strategies, richer haptic/visual parameterizations, and field evaluations that assess comprehension and user burden over time.

7 Future Work

A direct, head-to-head comparison of arousal-to-haptics (Part 1) and acoustic-to-haptics (Part 2) within a unified paradigm could clarify their respective strengths and trade-offs, including interpretability, cross-modal consistency, and cognitive load. Given the role of individual preference observed across both parts, future work could employ participatory and co-design approaches with DHH users to refine mappings that better align with users' mental models, contexts of use, and desired levels of control.

Longitudinal studies could examine how learning, habituation, and fatigue influence the perception and usefulness of affective haptic patterns over time, particularly in longer-form content, potentially incorporating parameter personalization as a mitigating strategy to explore. Finally, future systems might investigate alternative affective constructs—*e.g.,* valence, dominance, or affective relevance—alongside or instead of arousal to better capture the dimensions of speech that DHH viewers find meaningful.

8 Conclusion

Haptic feedback can enrich access to vocal affect for DHH viewers, especially in short-form, caption-heavy video where quick cuts, overlays, and on-screen text strain vision. In our study, arousal alone proved an unstable target: participants held diverse mental models and often mapped "more" vibration to loudness rather than emotion. Shifting from single-parameter arousal dials to multi-parameter acoustic → haptic mappings yielded clear emotion judgments (*e.g.,* PULSE for high arousal; PITCH-NORMALIZED for low). For creators and platforms, offloading affective cues to the wrist or hands—emphasizing *contrast* over absolute energy, combining rhythm/pitch/waveform, aligning tactile with visual cues, and preserving user control—can make short clips emotionally resonant without crowding the screen. These findings offer actionable guidance for haptic-infused captioning that conveys not just words, but also the *feeling* behind them.

Acknowledgments

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

The authors acknowledge the use of OpenAI's ChatGPT and Google's Gemini as editorial tools for grammar checking and improving textual clarity. All text was reviewed by the authors to ensure it accurately reflects their intended meaning.

We thank Yiwen Wang for sharing the STL files used to 3D-print the Vibro-Transducer casing, and the members of the Center for Accessibility and Inclusion Research (CAIR) lab at Rochester Institute of Technology for their help with reviewing the paper.

References

[1] Joshua M. Ackerman, Christopher C. Nocera, and John A. Bargh. 2010. Incidental Haptic Sensations Influence Social Judgments and Decisions. *Science* 328, 5986 (June 2010), 1712–1715. doi:10.1126/science.1189993

[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. doi:10.1145/2702123.2702288

[3] Lisa Feldman Barrett and James A. Russell. 1999. The Structure of Current Affect: Controversies and Emerging Consensus. *Current Directions in Psychological Science* 8, 1 (Feb. 1999), 10–14. doi:10.1111/1467-8721.00003

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

[5] 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.

[6] Federal Communications Commission. 2021. Closed Captioning on Television | Federal Communications Commission. https://www.fcc.gov/consumers/guides/closed-captioning-television

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

[8] Caluã de Lacerda Pataca, SooYeon Ahn, Suhyeon Yoo, JooYeong Kim, Khai N. Truong, Jin-Hyuk Hong, Roshan L. Peiris, and Matt Huenerfauth. 2025. CuCap: Comparative Analysis of Customized Captioning between North American and South Korean d/Deaf and Hard-of-Hearing Users. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25)* (Denver, CO, USA). Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3663547.3746400

[9] 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. doi:10.1109/TAFFC.2022.3174721

[10] 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. doi:10.1145/3544548.3581511

[11] Caluã de Lacerda Pataca, Saad Hassan, Lloyd May, Michelle M Olson, Toni D'aurio, Roshan L Peiris, and Matt Huenerfauth. 2025. Tactile Emotions: Multimodal Affective Captioning with Haptics Improves Narrative Engagement for d/Deaf and Hard-of-Hearing Viewers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 68, 17 pages. doi:10.1145/3706598.3713304

[12] 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. doi:10.1145/3613904.3642258

[13] Otniel E. Dror. 2024. The Feeling "Without Any Name". *Emotion Review* 17, 1 (Dec. 2024), 16–18. doi:10.1177/17540739241303492

[14] Paul Ekman and Wallace V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124–129. doi:10.1037/h0030377

[15] 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. doi:10.1145/3290605.3300276

[16] Mark D. Fletcher. 2021. Using haptic stimulation to enhance auditory perception in hearing-impaired listeners. *Expert Review of Medical Devices* 18, 1 (2021), 63–74. arXiv:https://doi.org/10.1080/17434440.2021.1863782 doi:10.1080/17434440.2021.1863782 PMID: 33372550.

[17] Mark D. Fletcher, Nour Thini, and Samuel W. Perry. 2020. Enhanced Pitch Discrimination for Cochlear Implant Users with a New Haptic Neuroprosthetic. *Scientific Reports* 10, 1 (June 2020), 10354. doi:10.1038/s41598-020-67140-0

[18] Addison Gage, Anthony Sahaf, Pranav Pidathala, Matthew Seita, and Christian Vogler. 2025. Haptic Feedback in Movies with a Sony PS5 Controller. *Journal on Technology and Persons with Disabilities* 13 (2025), 240–255.

[19] E. Gatti, G. Caruso, M. Bordegoni, and C. Spence. 2013. Can the feel of the haptic interaction modify a user's emotional state?. In *2013 World Haptics Conference (WHC)*. IEEE, Daejeon, South Korea, 247–252. doi:10.1109/whc.2013.6548416

[20] 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. doi:10.1145/642611.642653

[21] Morton Ann Gernsbacher. 2015. Video Captions Benefit Everyone. *Policy Insights from the Behavioral and Brain Sciences* 2, 1 (Oct. 2015), 195–202. doi:10.1177/2372732215602130

[22] Martijn Goudbeek and Klaus Scherer. 2010. Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America* 128, 3 (Sept. 2010), 1322–1336. doi:10.1121/1.3466853

[23] 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. doi:10.1145/3544549.3585880

[24] Hearing Loss Association of America. 2025. Hearing Loss by the Numbers. https://www.hearingloss.org/understanding-hearing-loss/hearing-loss-101/hearing-loss-by-the-numbers/

[25] Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71 (2018), 1–15. doi:10.1016/j.wocn.2018.07.001

[26] 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. doi:10.1145/3544548.3581130

[27] Peter Kuppens, Francis Tuerlinckx, James A. Russell, and Lisa Feldman Barrett. 2013. The relation between valence and arousal in subjective experience. *Psychological Bulletin* 139, 4 (July 2013), 917–940. doi:10.1037/a0030811

[28] Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham. 2013. Captions versus transcripts for online video content. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (Rio de Janeiro, Brazil) *(W4A '13)*. Association for Computing Machinery, New York, NY, USA, Article 32, 4 pages. doi:10.1145/2461121.2461142

[29] Steven R. Livingstone and Frank A. Russo. 2019. RAVDESS Emotional speech audio. doi:10.34740/KAGGLE/DSV/256618

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

[31] Lloyd May, Michael Clemens, Khang Dang, Keita Ohshiro, Sripathi Sridhar, Pauline Wee, Magdalena Fuentes, Sooyeon Lee, and Mark Cartwright. 2025. "Choices? That's the dream": challenges and opportunities in non-speech information closed-captioning. *Frontiers in Computer Science* 7, Article 2-s2.0-105009903599 (2025), 22 pages. doi:10.3389/fcomp.2025.1575176

[32] Lloyd May, Keita Ohshiro, Khang Dang, Sripathi Sridhar, Jhanvi Pai, Magdalena Fuentes, Sooyeon Lee, and Mark Cartwright. 2024. Unspoken Sound: Identifying Trends in Non-Speech Audio Captioning on YouTube. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 57, 19 pages. doi:10.1145/3613904.3642162

[33] Lloyd May, So Yeon Park, and Jonathan Berger. 2023. Enhancing Non-Speech Information Communicated in Closed Captioning Through Critical Design. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 16, 14 pages. doi:10.1145/3597638.3608398

[34] Antonella Mazzoni and Nick Bryan-Kinns. 2016. Mood Glove: A haptic wearable prototype system to enhance mood music in film. *Entertainment Computing* 17 (Nov. 2016), 9–17. doi:10.1016/j.entcom.2016.06.002

[35] Emma J McDonnell and Leah Findlater. 2024. Envisioning Collective Communication Access: A Theoretically-Grounded Review of Captioning Literature from 2013-2023. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 23, 18 pages. doi:10.1145/3663548.3675649

[36] 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. doi:10.1145/2331714.2331745

[37] 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.). Zenodo, Mexico City, Mexico, Article 28, 6 pages. doi:10.5281/zenodo.11189155
[38] Stephen Nixon, Lisa Huang, Katja Schimmel, Rafał Buchner, and Cris R Hernández. 2023. Recursive Sans & Mono. http://www.recursive.design/
[39] National Association of the Deaf. 2017. Closed Captioning Requirements. https://www.nad.org/resources/technology/television-and-closed-captioning/closed-captioning-requirements/
[40] Maja Pantic, Nicu Sebe, Jeffrey F. Cohn, and Thomas Huang. 2005. Affective multimodal human-computer interaction. In *Proceedings of the 13th Annual ACM International Conference on Multimedia* (Hilton, Singapore) *(MULTIMEDIA '05)*. Association for Computing Machinery, New York, NY, USA, 669–676. doi:10.1145/1101149.1101299
[41] Michael V. Perrotta, Thorhildur Asgeirsdottir, and David M. Eagleman. 2021. Deciphering Sounds Through Patterns of Vibration on the Skin. *Neuroscience* 458 (2021), 77–86. doi:10.1016/j.neuroscience.2021.01.008
[42] Rosalind W. Picard. 1997. *Affective Computing*. The MIT Press, Cambridge, MA, USA. doi:10.7551/mitpress/1140.001.0001
[43] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. doi:10.48550/ARXIV.2212.04356
[44] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.
[45] Katri Salminen, Veikko Surakka, Jani Lylykangas, Jussi Rantala, Teemu Ahmaniemi, Roope Raisamo, Dari Trendafilov, and Johan Kildal. 2012. Tactile Modulation of Emotional Speech Samples. *Advances in Human-Computer Interaction* 2012 (2012), 1–13. doi:10.1155/2012/741304
[46] David Sander. 2024. Is "Arousal," as a Scientific Concept, Worse than Useless? *Emotion Review* 17, 1 (Dec. 2024), 19–22. doi:10.1177/17540739241303501
[47] 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.
[48] 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. doi:10.1145/3544549.3585625
[49] 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. doi:10.1145/1978942.1979044
[50] Karen E. Smith, Kristina Woodard, and Seth D. Pollak. 2025. Arousal May Not Be Anything to Get Excited About. *Emotion Review* 17, 1 (2025), 3–15. arXiv:https://doi.org/10.1177/17540739241303499 doi:10.1177/17540739241303499
[51] Petra Sundström, Anna Ståhl, and Kristina Höök. 2007. In situ informants exploring an emotional mobile messaging system in their everyday practice. *International Journal of Human-Computer Studies* 65, 4 (April 2007), 388–403. doi:10.1016/j.ijhcs.2006.11.013
[52] Luca Turchet, Raffaele Rosaia, Alessandro Diodati, and Marco Carner. 2025. Exposure to vibrotactile music improves audiometric performances in individuals with cochlear implants. *Scientific Reports* 15, 1, Article 2-s2.0-105007928565 (June 2025), 11 pages. doi:10.1038/s41598-025-02946-4
[53] World Wide Web Consortium (W3C). 2018. *Web Content Accessibility Guidelines (WCAG) 2.1*. World Wide Web Consortium (W3C). https://www.w3.org/TR/WCAG21/
[54] 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. doi:10.1109/TPAMI.2023.3263585
[55] 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. arXiv:https://direct.mit.edu/comj/article-pdf/39/4/10/1953737/comj_a_00324.pdf doi:10.1162/COMJ_a_00324
[56] 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. doi:10.1145/3544548.3581076
[57] Christopher D. Wickens. 2008. Multiple Resources and Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, 3 (June 2008), 449–455. doi:10.1518/001872008x288394
[58] 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. doi:10.1080/15213260701283079
[59] 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.
[60] Nianmei Zhou, Steven Devleminck, and Luc Geurts. 2024. Tangible Affect: A Literature Review of Tangible Interactive Systems Addressing Human Core Affect, Emotions and Moods. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 424–440. doi:10.1145/3643834.3661608
[61] Elvar Atli Ævarsson, Thórhildur Ásgeirsdóttir, Finnur Pind, Árni Kristjánsson, and Runar Unnthorsson. 2022. Vibrotactile Threshold Measurements at the Wrist Using Parallel Vibration Actuators. *ACM Transactions on Applied Perception* 19, 3 (July 2022), 1–11. doi:10.1145/3529259

A Part 1: Arousal-to-Haptic Pulse Mapping

In Part 1, each caption cue is annotated with an *arousal* value $a \in [0, 1]$ and a cue duration d (in seconds). When haptics are enabled, we render a *single short vibration* on the wrist for each caption, using a fixed carrier frequency of 75 Hz and an amplitude proportional to arousal.

Concretely, for each caption cue we compute a target amplitude $A \in [A_{\min}, A_{\max}]$ from a, and drive a 75 Hz sinusoid through an attack–sustain–release envelope that spans the cue duration d. The implementation is realized in WebChucK by generating a short ChucK script per cue, but the logic can be summarized as Algorithm 1.

Algorithm 1: Part 1: Single-Pulse Arousal-to-Haptics Mapping (75 Hz)

Input: caption cue with arousal $a \in [0, 1]$ and duration $d > 0$
Output: control parameters for a single haptic pulse on the wrist
Data: fixed carrier frequency $F = 75$ Hz, minimum amplitude $A_{\min}$
`// 1. Map arousal a to an output amplitude A`
$A \leftarrow \max(a, A_{\min})$
`// 2. Define a simple attack–sustain–release envelope over duration d`
`// Split the cue into three equal phases`
$T_{\text{seg}} \leftarrow d/3$
$T_{\text{attack}} \leftarrow T_{\text{seg}}$
$T_{\text{sustain}} \leftarrow T_{\text{seg}}$
$T_{\text{release}} \leftarrow T_{\text{seg}}$
`// 3. Synthesis (realized in ChucK/WebChucK)`
Create a 75 Hz sine oscillator and envelope:
(1) Set oscillator frequency to F.
(2) Set envelope target to A with duration T_{attack}, then trigger key-on.
(3) Wait $T_{\text{attack}} + T_{\text{sustain}}$ seconds.
(4) Set envelope duration to T_{release} and trigger key-off.
(5) Wait T_{release} seconds to complete the pulse.

B Part 2: Speech-to-Haptics Mappings

This appendix provides detailed pseudocode for all acoustic-to-haptic mappings used in Part 2 of the study. All mappings share a common frame-level analysis stage: the audio signal is segmented into short windows, and for each window we extract two acoustic parameters:

- **Pitch** $F0_f$ (in Hz): fundamental frequency estimated using Parselmouth. When a frame is unvoiced, pitch is undefined and the frame is discarded.
- **Intensity** I_f (dB): averaged intensity within the frame. This is used to detect and ignore silence within the frame, to avoid sending noisy pitch information.

All five mappings therefore take as input a stream of $F0_f$ data and transform it into haptic control parameters. Each mapping specifies:

- how pitch is modified (*e.g.,* clipped, rescaled, exaggerated),
- how amplitude is applied (continuous or pulsed),
- and, in one case, which waveform is used (*sine* vs. *sawtooth*).

The Python implementation transmits the pitch parameter via OSC messages (`/pitch`). On the receiving end, a ChucK synthesizer updates oscillator frequency and amplitude in real time. Algorithm 2 describes the shared analysis stage, followed by Algorithms 3–7 which define each mapping. Algorithm 2 specifies the OSC-to-synthesis loop used in all conditions.

Algorithm 2: Part 2: Frame-Level Acoustic Analysis (common to all mappings)

Input: audio waveform $x(t)$
Output: sequence of pitch data $F0_f$
Data: window size w, hop size h, intensity threshold $I_{\min}$
for *each frame* f *in* $x(t)$ *with step* h **do**
  compute average pitch for window $F0_f$
  compute average intensity for window I_f
  if $F0_f < 80$ *or* $F0_f > 300$ **then**
    skip frame (pitch outside of speech range)
  if $I_f < I_{\min}$ **then**
    skip frame (treat as silence)
  else
    send $F0_f$ to the selected haptic mapping

Algorithm 3: Part 2: Unfiltered Mapping (baseline)

Input: sequence of pitch data $F0_f$
Output: OSC messages to haptic synthesizer
for *each frame* f **do**
  $f_{\text{haptic}} \leftarrow F0_f$
  send OSC message
    `/pitch` = f_{haptic}

Algorithm 4: Part 2: Pitch-Normalized Mapping

Input: sequence of pitch data $F0_f$
Output: OSC messages to haptic synthesizer
Data: vibrotactile band $[F_{\min}, F_{\max}]$
for *each frame* f **do**
  $f_{\text{haptic}} \leftarrow \text{clip}(F0_f, F_{\min}, F_{\max})$
  send OSC message
    `/pitch` = f_{haptic}

Algorithm 5: Part 2: Pulse Mapping

Input: sequence of pitch data $F0_f$
Output: OSC messages to haptic synthesizer
Data: pulse on-duration T_{on}, pulse off-duration T_{off}
for *each frame* f **do**
  $f_{\text{haptic}} \leftarrow \text{clip}(F0_f, F_{\min}, F_{\max})$
  `// Emit a short vibration pulse followed by silence`
  send OSC message
    `/pitch` = f_{haptic}
  wait until T_{off} has elapsed to continue

Algorithm 6: Part 2: Sawtooth Mapping

Input: sequence of pitch data $F0_f$
Output: OSC messages to haptic synthesizer
Data: vibrotactile band $[F_{\min}, F_{\max}]$
for *each frame* f **do**
  $f_{\text{haptic}} \leftarrow \text{clip}(F0_f, F_{\min}, F_{\max})$
  send OSC message with waveform flag
    `/waveform` = "sawtooth"
  send OSC message
    `/pitch` = f_{haptic}

Algorithm 7: Part 2: Pitch-Exaggerated Mapping

Input: sequence of pitch data $F0_f$
Output: OSC messages to haptic synthesizer
Data: breakpoint $B = 120$ Hz, scale factor $\alpha > 1$, band $[F_{\min}, F_{\max}]$
for *each frame* f **do**
  $f_{\text{haptic}} \leftarrow \text{clip}(f_{\text{haptic}}, F_{\min}, F_{\max})$
  if $f_{\text{haptic}} > B$ **then**
    $f_{\text{haptic}} \leftarrow B * \alpha$
  send OSC message
    `/pitch` = f_{haptic}