Haptic-Captioning: Using Audio-Haptic Interfaces to Enhance Speaker Indication in Real-Time Captions for Deaf and Hard-of-Hearing Viewers

Yiwen Wang*
yw7615@umd.edu
College of Information Studies,
University of Maryland
College Park, Maryland, USA

Ziming Li
zl1398@rit.edu
School of Information,
Rochester Institute of Technology
Rochester, New York, USA

Pratheep Kumar
pc9099@rit.edu
School of Information,
Rochester Institute of Technology
Rochester, New York, USA

Wendy Dannels
w.dannels@rit.edu
Center on Culture and Language,
National Technical Institute for the Deaf,
Rochester Institute of Technology
Rochester, New York, USA

Tae Oh
tae.oh@rit.edu
School of Information,
Rochester Institute of Technology
Rochester, New York, USA

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

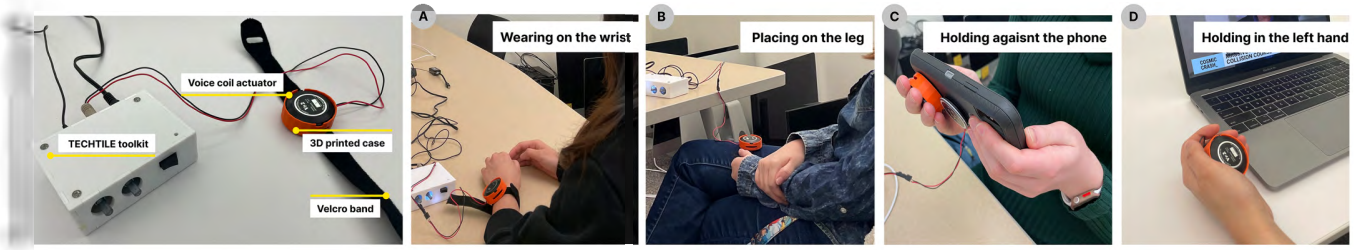

Figure 1: Left-The Haptic-Captioning system consists of a voice-coil actuator, a power amplifier (TECHTILE Toolkit [30]), a 3D printed case, and a velcro band; Right (A)-(D): Four vibration positions that participants preferred to use: wearing on the wrist, placing on the leg, holding against the phone, holding in the right hand.

ABSTRACT

Captions make the audio content of videos accessible and understandable for deaf or hard-of-hearing people (DHH). However, in real-time captioning scenarios, captions alone can be challenging for DHH users to identify the active speaker in a real time in multiple-speaker scenarios. To enhance the accessibility of real-time captioning, we propose Haptic-Captioning which provides real-time vibration feedback on the wrist by directly translating the sound of content into vibrations. We conducted three experiments to examine: (1) the haptic perception (Preliminary Study), (2) the feasibility of the haptic modality along with real-time and non-real-time visual captioning methods (Study 1), and (3) the user experience of using the Haptic-Captioning system in different media contexts (Study 2). Our results highlight that the Haptic-Captioning complements visual captions by improving caption readability, maintaining media engagement, enhancing understanding of emotions, and assisting speaker indication in real-time captioning scenarios. Furthermore, we discuss design implications for the future development of Haptic-Captioning.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility technologies**.

KEYWORDS

Haptics, captioning, accessibility, deaf and hard of hearing

ACM Reference Format:
Yiwen Wang, Ziming Li, Pratheep Kumar, 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 (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3544548.3581076

*This work while affiliated with Rochester Institute of Technology.

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

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581076

1 INTRODUCTION

Captions are widely used in different media sources to support deaf and hard-of-hearing (DHH) viewers to follow the aural-based dialogues. However, reading captions alone could be insufficient, especially when there are multiple speakers in a scene such as a live discussion with multiple panelists or a live event with multiple commentators on television (TV). In such unscripted situations where *real-time captions* are presented through captioners or auto-generated through Automatic Speech Recognition (ASR), a typical challenge is identifying and indicating the speaker in the captions when the conversations could rapidly switch between multiple speakers [22]. Moreover, research has indicated that this could be a tiring task for DHH individuals when having to switch between viewing captions and identifying the speakers frequently [14, 22].

To improve the caption accessibility, most prior studies have been focused on speaker indication through visual design (e.g., using different colors [3], placing the text under the speaker [18], inserting speaker names [36], adding the avatar image of speakers [36], highlighting the active speaker through pop-ups [22], and signifiers [14]). However, these non-real-time captioning methods may be challenging for real-time captioning as the captioning methods' performance (such as ASR) is limited in identifying speakers in multi-speaker environments [5] or it can cause significant delays to display the captions with speaker identifications [1, 3, 28]. In addition, concerns have been discussed about the extra cognitive loads required with visual speakers indication schemes such as recalling the visual cues for each speaker [1] and, signifiers and text with constantly changing positions would be distracting on video watching [22].

To address the challenges above, we propose *Haptic-Captioning*, an audio-haptic based system to enhance the traditional real-time captioning for live media content (e.g., news, sports) (Figure 1). This multi-modal system aims to support DHH viewers on speaker indication by advancing the understanding of paralinguistic aspects of communication. The Haptic-Captioning system directly translates the auditory content into vibrations using a voice-coil, which is a haptic actuator that is powered by an audio power-amplifier [30]. As such, this method generates vibrations that carry the same properties as the audio signal, preserving characteristics such as the loudness and pitch of the voices. Therefore, when presented along with the traditional captions, we posit that DHH users can identify speakers with the Haptic-Captioning system, similar to how hearing individuals identify different speakers by the unique characteristics of the voices. Similar wrist-worn audio-haptic systems have been frequently utilized to provide sound awareness for DHH users based on the characteristics of the sound [12, 15]. Audio-haptic systems have also been explored towards enhancing captions, specifically, to present non-speech information (NSI) such as an object falling or a phone ringing in a movie scene [23].

In this research, we explored the Haptic-Captioning system as a wearable, holdable or attachable device for providing enhanced haptic feedback for captions (Figure 1). To evaluate our system, we conducted a Preliminary Study followed by two user studies with a total of 34 DHH participants. Our Preliminary Study aimed at gaining initial insights into the speaker indication capabilities just using only haptic feedback with 12 DHH participants. The participants were asked to wear the Haptic-Captioning device on their wrists and discuss the perceived number of speakers and other perceived information by observing the audio clips via only vibrations. Next, in Study 1, we conducted a comparative study with 16 DHH participants to examine speaker indication accuracy and user preference between the Haptic-Captioning system and other real-time and non-real-time visual speaker indication methods. Motivated by the results in the 2 previous studies, in Study 2, we focused on a qualitative approach with a contextual interview by providing users with different genres of content (i.e., podcast, sports, live stream, movie) and settings (i.e., mobile, TV, laptop) together with the Haptic-Captioning system. While this system can possibly be adopted in the real-time conversation between multiple speakers, our work mainly focuses on the media experience for content that is being captioned live.

In summary, our main contributions include: (1) The Haptic-Captioning system that directly translates audio into haptic patterns to enhance real-time captions. (2) A preliminary study of the Haptic-Captioning system's speaker indication characteristics with DHH participants. (3) A comparison and discussion on speaker indication accuracies and the user preferences of haptic, real-time, and non-real-time captioning methods. (4) A contextual study on DHH people's experience of Haptic-Captioning in different contexts of use and design implication for future Haptic-Captioning devices.

2 RELATED WORK

2.1 Speaker Indication Accessibility and Challenges

Previous research on interactive television and media experiences had limited consideration of accessibility (4.29%) and inclusion of disabled participants (2%) in the study out of 449 publications happened between 2003 and 2020 [35]. With respect to the viewing experience of media content, previous work stressed the importance of video accessibility through captions which proves to improve attention and comprehension for people learning to read, understanding non-native languages, and for DHH people [13]. To enhance the accessibility of video content for DHH people, Butler explored how captions can influence the viewing experience [8]. Qualitative analysis of the study discusses the viewing balance of captions, and the visual design of captions such as font, color, background, size, and length of lines. The previous study recommends processing different aesthetic and accessible designs for captions based on individual preference and engagement with the visual-aural content.

Another research on improving caption accessibility examined the preference of caption positions to avoid occlusion in videos having text-rich content [2]. Their findings contributed to defining guidelines for caption placement and caption-evaluation methods for live television genres. There are many opportunities to improve caption accessibility where a study by Vy and Fels, underlined the difficulty in identifying speaker change for media content that has multiple speakers, narrative discussions, and off-screen speakers [36].

2.2 Existing Speaker Indication Methods

Prior works on speaker indication in a captioned video focused on three aspects: caption positioning, visual cue indication, and textual cue embedded with the caption. Placing captions dynamically closer to the speaker in the video utilizes facial recognition aspects such as motion region prediction and lip movements to determine the speaker in the video content [7, 17, 18, 22, 29]. This will help reduce the disconnection between the visual location of the speaker in the video and the caption area. However, it can result in visual overload when there are overlapping multiple speakers present in a single scene and can cause eyestrain following up between each dynamically shifting position of captions and speaker indicators. A technique that uses visual cues such as lightbulb, glow, and pointing methods to indicate the current speaker addressed this dynamic caption position shift, especially in a panel presentation with unpredictable switching among multiple speakers, and maintained a separate single static position to display the captions [14]. Researchers implemented this technique in a head-mounted display to conduct the study with DHH participants and suggested it be easier in identifying the speakers [14, 20]. But these techniques might not be efficient to identify the speaker when they are not visible in the scene while speaking. Another visual cue method to indicate the current speaker used avatar badge with the name, character image, and colored border of the character's cloth color [36]. But the study was reported to be distracting and less useful as DHH participants prefer to identify the speaker by the physical appearance and personality rather than the speaker's name in a video.

2.3 Audio-Haptic Methods

Auditory perception also helps people to familiarize a particular voice to identify a speaker based on the time, frequency, intensity, and pitch of the sound waves [32]. Audio-haptic technologies have been widely used in recent research for a wide range of applications [9, 10, 25–27, 30]. Among these, many works have explored using audio-haptic or sound-based haptic to make everyday sounds accessible for DHH users [31, 34]. Study conducted by Weisenberger et al., [40] used two 16-channel tactile feedback devices worn on the forearm and abdomen for phoneme discrimination task. This multichannel Tactile feedback setup has proved to provide a better perception of speech, especially when combined with lipreading [40] and also performed better than the single-channel tactile system for the majority of speech perception task [39]. To enhance the sound awareness of DHH people, tactile technology has been used to identify sound patterns through a wrist-worn device that emits haptic feedback based on the sound level. [12, 19]. This study [19] being one of the primary motivations behind our work suggests that vibrotactile information enhances the sound "experience" in the environment through an evaluation in a life field experiment.

Research has also been done on enhancing caption accessibility through visual-tactile information [23]. Here, Kushalnagar et al. conducted a study to enhance the caption experience for non-speech information by presenting visual-tactile captions and suggested an increase in viewing and recall ability for DHH people. Another study explored experience tactile technology for the entire human body through chairs [37] where haptic sensory was placed on various places such as armrest, back-rest [31], and under the seat [38].

2.4 Summary and Research Questions

These works helped us understand the accessibility challenges to identifying the speaker changes in various video content including live videos where the captions cannot be pre-processed. Building on previous work, there are three research questions we would like to answer in this study. RQ1: How do DHH viewers perceive speaker information in the media content through haptic feedback *alone*? RQ2: What are the user preferences and efficiency of speaker indication of Haptic-Captioning modality compared to existing real-time and non-real-time captioning methods? RQ3: How does Haptic-Captioning system affect DHH's user experience and what factors should be considered in future designs?

3 HAPTIC-CAPTIONING SYSTEM DESIGN

The Haptic-Captioning system, shown in Figure 1-Left, produces haptic vibrations through *voice coil actuators*. These are vibrotactile devices that vibrate using sound signals, similar to an audio speaker without a cone to amplify sound — therefore, voice-coil actuators may emit a slight sound when in use. In this research, we use the Acouve Vp210[1] as our voice coil actuator. To actuate it, we use a power amplifier based on the Techtile Toolkit design [30] (any power amp up to 3W can be used to drive this actuator). Thus, any audio signal from any input source such as a laptop, phone, etc., can be used to drive the voice coil actuator. In addition, the intensity of the vibrations can be changed by adjusting the volume on the input source and/or the power amplifier.

We designed and 3D printed a casing for the actuator and used a velcro band similar to a wristband as previous research demonstrated the haptic sensitivity of the skin on the wrist [10, 12, 19]. We evaluated the Haptic-Captioning system through a three-phase experiment where we used the wristband prototype for the Preliminary Study and Study 1. For Study 2, we explored other possible uses of the Haptic-Captioning system such as while held or attached to a device such as a mobile phone based on the requirement.

4 PRELIMINARY STUDY OF THE HAPTIC-CAPTIONING SYSTEM

As the Preliminary Study[2], we aimed to qualitatively explore the DHH users' ability to perceive speaker characteristics such as the number of speakers through only haptic feedback (RQ1) without any visual or auditory feedback. To achieve this goal, we recruited twelve DHH volunteers (5 females, 6 males, and 1 non-binary) aged from 18 to 44 (M=26.3, SD=7.3) for this study. Seven of them identified themselves with profound hearing loss, and five of them identified themselves with mild or severe hearing loss.

4.1 Study Design

For this Preliminary Study, we selected sixteen 1-minute audio clips from live stream videos on YouTube[3] which involved topics

[1] https://www.acouve-lab.com/products

[2] All studies listed in this paper were approved by the Ethics Board of the Institution. In all studies, each participant was paid $15 for their participation

[3] https://www.youtube.com

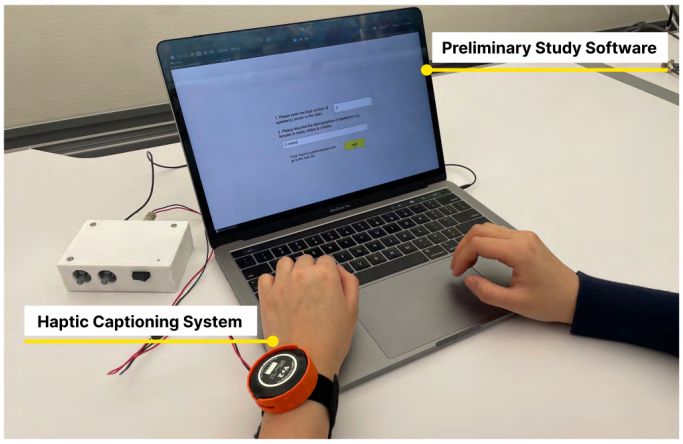

Figure 2: The Preliminary Study set-up consists of the Haptic-Captioning system that provides auditory-haptic feedback and the study software which is used to collect the perceived speaker information.

such as sustainability, life wisdom, business, education, and fashion. We excluded captions and visuals from the videos as we wanted the participants to *only focus on the haptic feedback* and avoid being biased by the content of the captions or visual cues, for example, the video zooming on the speakers or the lip reading. The clips evaluated in this Preliminary Study consisted of two to three speakers and different speakers' demographics with multiple age groups, perceived genders, etc.

4.2 Study Procedure

In the study session, each participant was expected to experience the Haptic-Captioning system through 16 trials corresponding to 16 different audio clips. Prior to the study, we set up the wearable vibrator on the participant's wrist. Participants also wore a pair of headphones that played white noise to avoid the participant hearing any sound leaks from the voice coil actuator as we only wanted to focus on the effects of the haptic feedback. For each round, we played a random audio clip with haptic feedback to the participant. Then, we asked the participant to discuss the perceived number of speakers and other perceived information (e.g., perceived gender, age, etc.) (Figure 2). After the study, we conducted an interview to understand participant's overall experience and other information such as how they perceived the different speakers, perceived speakers' demographics, etc., through the haptic patterns. It took approximately 45 minutes for each participant to complete a study session.

4.3 Findings of Preliminary Study

Overall, the participants identified the number of speakers with a 72.34% accuracy (two speakers: $M = 79.168\%$, three speakers: $M = 70.921\%$). Feedback revealed more insight into their methods for using vibrations during this task.

Participants indicated that they tried to identify speaker demographics and thereby the number of speakers by understanding the vibration intensity to infer the speaker's demographic characteristics. For example, P08 noted that, *"If the haptic feedback seemed deeper/louder, I thought it might be a man. If it was lighter/quieter, I thought it was a woman."* Except for the perceived gender of the speakers, participants also used the vibration strength to infer the speaker's age group. Here, P09 stated that, *"Heavy buzz or high buzz makes me think of adult and male... high light buzz makes me think of female. Low softer buzz makes me think of children. Sometimes it messes up [when there is] a combination of buzzes of either softer buzz and high pitch which could be a female speaker [who is] loud."*

In terms of identifying the number of speakers, the participants, such as P02, indicated that they can perceive the amount of speakers through the differences of speaking voices conveyed by the vibration: *"I can notice few speakers at least 3-4 speakers. I can tell the difference between the deep and soft voice, but not sure which genders/ages they belong to."* Two participants (P04 & P12) stated that they surmised the transition of speakers by comparing the vibration pattern of the speaker with the previous one after each pause and to infer the speaker's amount according to this evidence. As it was mentioned by P12, *"The vibrations change when the speakers switched — I would notice the vibration would either stop for a bit or have a new kind of vibration that was different. So I assume they switch speakers."*

Participants also expressed their concerns on the testing scenarios with multiple speakers, especially when the speakers had similar voice patterns. For example, P05 mentioned that, *"It was challenging to try and identify multiple different speakers. If two speakers have a similar tone, I would not be able to recognize that. I had to second-guess myself at times and really assess whether or not I was feeling a difference in vibrations."*

4.4 Summary of the Preliminary Study

In this Preliminary Study, we identified that the mean accuracy of the speaker's amount predictions was 72.34%, which was a promising accuracy rate considering the system evaluated in the study only used vibration feedback. It should be noted that this study consisted of the *only* vibration feedback without any usual visual (video content or captions) or auditory feedback. Thus, the preliminary study revealed that the vibration feedback of the Haptic-Captioning system can convey perceptual knowledge to a certain extent, such as

the number of speakers and other speakers' demographics, to the DHH participants. However, the participants might still feel uncertain of their answers when factors like speakers' tone changed or similar voice patterns of the speakers were involved. Therefore, the next step of our study is to explore if the overall performance or user experience of the Haptic-Captioning system at speaker indications will be improved if combining other cues like visuals or captions with haptics as a multimodal feedback system.

5 STUDY 1: CAPTION METHODS COMPARISON

Our Preliminary Study indicated that the audio-generated haptic feedback was promising at identifying various speaker characteristics. Inspired by the participants' feedback collected from the Preliminary Study, we wanted to further explore the usability of the Haptic-Captioning system when combined with other cues such as visuals and captions. In addition, we compared it with prior non-real-time visual methods of speaker identification for captions [3, 11, 14, 36] as well as the real-time captioning methods (RQ2). As the dependent variables, we included Perceived Speaker Transitions and Self-Reported Questionnaires followed by discussions with the participants.

5.1 Study Design

To answer RQ2, we designed a within-subjects evaluation for the comparative study that consisted of the *Caption Modality* as the main independent variable. The Caption Modality consisted of seven conditions: the proposed Haptic-Captioning method, the six visual captioning styles shown in Figure 3 (i.e. avatar, color, position, pointer, speaker name, and the traditional real-time captions as the baseline condition). The Haptic-Captioning condition used the traditional caption style in combination with the system but used with a different video.

5.1.1 Caption Modality. We selected several visual captioning methods based on previous research that can be considered as non-real-time methods as they are pre-processed videos with the caption styles added prior to the study. This is because, in real-time captioning contexts, visual captions may not be feasible as more information is needed to be identified to present the caption (identifying the position of the speaker, selecting color, etc.) and often, they have significant delays [3] that could change the viewer's experience. In contrast, the Haptic-Captioning system directly translates the audio into haptic vibrations. Following are the selected real-time and non-real-time caption methods.

Non-real-time captioning methods: *Avatar Caption* in Figure 3 (a) presents an image of a speaker with a distinguished color outline placed on the left side of plain text [36]. *Color Caption* in Figure 3 (b) is a color-coded method that uses different text colors to represent different speakers [3]. *Position Caption* in Figure 3 (c) places the text directly under the speaker and assists the speaker indication by changing the position. *Pointer Caption* in Figure 3 (d) uses a turn-on bulb to signify the active speaker while turn-off bubs represent speakers in silence [14]. *Speaker-name Caption* in Figure 3 (e) briefly presents the demographic speaker at the beginning of the sentence, e.g., Female Speaker 1 [11].

Real-time captioning methods: *Haptic-Captioning* method provides the tactile feedback generated from the auditory content with the *Traditional Real-time Caption* as shown in Figure 3 (f). The visual content were selected from a data set used in a previous study [3] and was presented as 30s videos with the corresponding captioning method added before the study. The order of caption modality conditions were randomized. In total, each participant tested seven trials.

5.1.2 Perceived Speaker Transitions. Previous work used user ratings to evaluate user preferences and experiences of different captioning methods [22, 23, 36]. Therefore, inspired by the participants' feedback in the preliminary study, to *quantitatively* evaluate the efficacy of the speaker indication capabilities of captioning methods, we propose a method that focuses on *speaker transitions*. Speaker transitions are defined as the number of times the speakers switched in a presented content. For example, when the first speaker asks a question and the second speaker answers, this is considered as a one-speaker transition. A participant may use the presented information from the different channels (visual, audio, haptic, etc.) to identify the speaker and the speaker transitions. Thus, using this method, a participant may report speaker transitions even in a situation where the speakers are not visible to identify (e.g., audio-podcast, commentators in a sports broadcast, etc.). In this scenario, some captioning methods such as colored captions could assist in indicating the speaker, while the methods such as position captions might not fulfill the task. While the concerns about the false alarms remain, we did not find any comments indicating participants were over-sensitive to report the speaker transition in the haptic condition. We coded all the visuals presented in this study to identify speaker transitions. This includes the number of speaker transitions and the time at which the speaker transition occurred. To analyze this data, we compared the time at which the participant reported a perceived speaker indication and summed all the correctly perceived speaker transition events. The accuracy is defined as the percentage of correctly perceived speaker transition events over the actual number of events (from the coded data).

5.1.3 Self-reported Questionnaire. Participants were asked to complete a 5-point Likert scale questionnaire after each trial. This questionnaire is designed with the goal of investigating the usability and caption effectiveness through the engagement of video content, distraction, comfortableness, fatigue, and difficulty of following the captions, caption understanding, and confidence on speaker indication [22]. We designed this subjective questionnaire to help us understand the user preferences and experience of the Haptic-Captioning method as well as other real-time and non-real-time captioning methods.

5.2 Apparatus

For this study, we used the same wearable prototype that was used in the Preliminary Study (Figure 4-Left). In the experiment software (developed using C# in the Unity platform), we presented the participant the video with the selected captioned conditions. Here, we also added a button right next to the video to click whenever the participant identified a speaker transition (Figure 4-Right). The procedure will be discussed more in the following sections.

Figure 3: Non-real-time Caption modalities: (a) Avatar Caption, (b) Color Caption, (c) Position Caption, (d) Pointer Caption, (e) Speaker-name Caption. Real-time Caption modality: (f) Traditional Real-time Caption - the same style but a different video used with the Haptic-Captioning.

Figure 4: (Left) Study 1 set-up consists of the Haptic-Captioning system that only turns on in the Haptic-Captioning condition and (Right) the screenshot of the Study 1 software interface to log the "speaker transition" data

5.3 Study Procedure

Firstly, we introduced the Haptic-Captioning system, the study procedure, and the study software to participants. The participants then received a demonstration of seven videos with different captioning styles in a knowing order. Each video was about 30 seconds. The introduction and demonstration allowed participants to familiarize themselves with the task. Participants could also adjust the intensity of haptic feedback during the introduction. We turned off the volume to avoid the potential effects resulting from different levels of hearing. Next, we asked participants to wear the haptic vibrator on their wrist. One researcher monitored the progress and only turned on the haptic device for the Haptic-Captioning condition. As the task, a participant was randomly presented with seven 30-second videos. Next, they used their other hand to click a "Mark" button via the touch pad when they identified a speaker transition. After each condition, the participants completed a self-reported questionnaire in Google Form to report their subjective experiences. Last, we asked for participants' preferences, challenges encountered, suggestions for the different contexts of use, overall experience using Haptic-Captioning as well as other real-time and non-real-time captioning methods through a semi-structured interview. The experiment took approximately 50 minutes per participant to complete.

ID	Gender	Hearing loss status	Lip reading familiarity	Hearing devices
P1	Male	Profound	Slightly	None
P2	Male	Severe	Extremely	Hearing aid/s
P3	Male	Mild	Slightly	Hearing aid/s
P4	Male	Profound	Extremely	Both
P5	Female	Profound	Not at all	None
P6	Female	Profound	Slightly	None
P7	Male	Profound	Not at all	None
P8	Male	Profound	Slightly	Cochlear implant/s
P9	Female	Mild	Moderately	None
P10	Female	Severe	Extremely	Hearing aid/s
P11/R1	Male	Profound	Extremely	Cochlear implant/s
P12/R2	Male	Mild	Somewhat	Hearing aid/s
P13/R3	Female	Moderate	Extremely	Hearing aid/s
P14/R4	Male	Profound	Somewhat	Hearing aid/s
P15/R5	Non-binary	Profound	Moderately	Cochlear implant/s
P16/R6	Female	Profound	Not at all	None

Table 1: Participants' demographic information

5.4 Participants

We recruited sixteen participants (P1-P16) as shown in Table 1 (nine males, six females, one non-binary) aged 18-35 (M = 23.8, SD = 4.7) from social media and the institute's mailing lists. As reported by the participants, ten of them had profound hearing loss, three had mild hearing loss, two had severe hearing loss, and one had moderate hearing loss. As for the hearing devices used, six participants used hearing aid(s), three used Cochlear implant(s), one used both hearing aid(s) and Cochlear implant(s), and six used none of the hearing devices. For the level of experience of lip-reading, five participants reported being extremely familiar, eight participants were at least slightly familiar, and three participants were not familiar at all. The detailed age information in Table 1 was removed to ensure the anonymity of participants.

5.5 Results and Discussion of Study 1

5.5.1 Perceived Speaker Transition. Figure 5 shows the overall results of the Perceived Speaker Transition accuracy. Overall, Haptic Caption scored the highest level of average accuracy with 93.75% (SD = 25). Lowest average accuracy of the non-real-time category was reported by *Pointer Caption* which was 73.75% (SD = 17.46), where as *Traditional Real-time Caption* scored 80.19% (SD = 20.36) average accuracy levels. We performed repeated measures one-way ANOVA on the accuracy of perceived speaker transition and found no significant differences ($F(6, 105) = 0.955, p = 0.459$).

Overall, the Haptic-Captioning condition achieved relatively high accuracy for detecting speaker transitions. Although we expected the non-real-time captions to perform better as participants have more previous experience using visual captioning than haptic ones, some participants indicated that it might be due to the nature of the visual content of the Haptic-Captioning condition's video that had only two speakers and no over-lapping conversations. While this was unintentional (we randomly chose the videos for each captioning type in the study design from the data set in [2, 3]), this brought our focus to the usability of the Haptic-Captioning system if there were frequent overlapping conversations. P4 commented on this aspect of overlapping conversations: *"Haptic-Captioning is very useful in identifying the switch in speakers, especially if the number of participants is few and the sound characteristics of the speakers are different. If the number of speakers is more and there is overlap in conversations, determining the switch in conversation becomes more difficult."*

Furthermore, we observed more mistakes that resulted in lower accuracies than expected for the visual captioning methods that could be due to the cognitive load required [36]. As for the visual captioning methods, P3 mentioned: *"It is putting more effort to move my eyes and effectively tell who is speaking depending on which caption methods. These challenges prevent me from fully immersed in the video content."* Participants also indicated that there were several challenges of the visual modalities. Participants mentioned the visual add-ons caused distraction and overwhelmed the reader. As P4 mentioned, *"associating the speaker information with visual cues cause higher cognitive load."* Similar observations also have been noted in previous works that investigated visual captioning styles [36]. Furthermore, P9 indicated: *"not all of the colors are friendly to use for those who are color blind."*

5.5.2 Self-reported questionnaire result. We conducted repeated measures one-way ANOVA to compare the Haptic-Captioning with the Traditional Real-time Caption, and five non-real-time captioning methods. The result is shown in Figure 6. Through the post-hoc pairwise comparison with the Holm correction, we identified there are significant differences between the rating of the Haptic-Captioning and the Position Caption (one of the non-real-time captioning methods) in the engagement ($p < 0.01$), confidence ($p < 0.001$), and comfortableness dimensions ($p < 0.01$). (For the scope of this work, we only report the significant effects between other caption methods against the Haptic-Captioning method.)

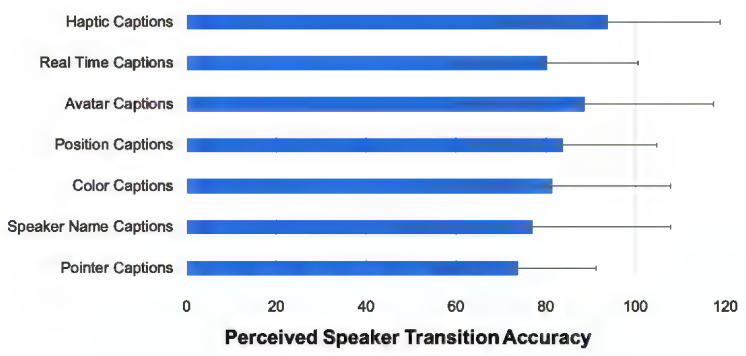

Figure 5: The average accuracy (percentage) of the perceived speaker transition. Error bars denote the standard deviation.

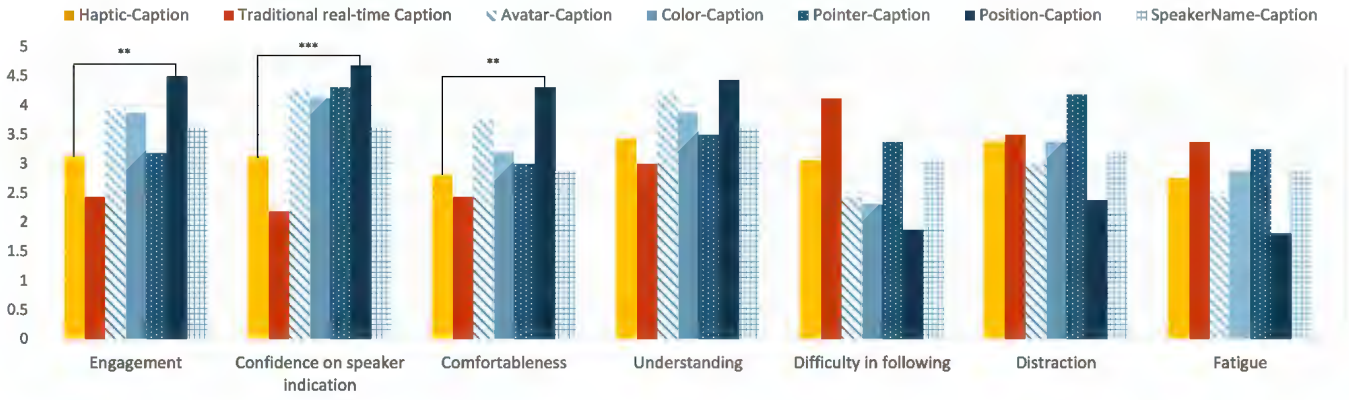

Figure 6: Self-reported ratings with the significant pairs (*: p<.001, **: p<.01, *: p<.05). For the Difficulty in following, Distraction, and Fatigue dimensions, a higher rating reflects a more negative experience. For the other dimensions, a higher rating reflects a more positive experience.**

In the results of the self-reported questionnaire, we identified that the Haptic-Captioning had better ratings compared with the Traditional Real-time Caption in all dimensions, including the Engagement of video content, Distraction, Understanding, Difficulty in following caption, Confidence on speaker identification, Comfortableness, and Fatigue. While comparing the Haptic-Captioning with the five non-real-time captioning methods, we found that the Haptic-Captioning did not significantly outperform in the comparisons. Especially in three dimensions: Engagement, Confidence on speaker identification, and Comfortableness, we noticed that the Position Caption, which was one of the non-real-time captioning methods, had significant better ratings compared to the Haptic-Captioning. Through the observation, the non-real-time captioning methods were more acceptable among the participants, since the methods that used visual cues for indications might possess a smaller learning curve. However, it should be noted that the participants also indicated the Haptic-Captioning system could be combined with the other captioning methods to utilize its strength, as reflected in the post-study interview result.

5.5.3 Post-study interview. We conducted a semi-structured interview to understand the characteristics of the Haptic-Captioning and how haptic feedback can be integrated into real-time captions. In analyzing the feedback, we found several themes.

Benefits of Haptic-Captioning system: Several participants found the Haptic-Captioning system to be useful. For example, P7 commented that haptic feedback is better than other modalities since they could feel a pause in the vibrations (in some cases) when there was switch in the speakers: *"I think the Haptic-Captioning is better than others so it helps me hear the vibration and read the caption in which it will switch the speaker so it looks like paused voice by the speaker."* P2 (who also tried holding the device besides wearing it on the wrist) discussed others way of using the system that could be beneficial: *"I was able to feel more with my fingers, but I can definitely see it becoming a thing that aides us. Possibly a dual paired device - or even something to add onto our seat of choice."*

Concerns about Haptic-Captioning system While we received some positive comments on Haptic-Captioning system, many participants expressed their concerns about wearing this device for a longer time. As P5 mentioned: *"I am not sure because if I have to wear it (the Haptic-Captioning system) for a long time. I will likely have sensory overload. "* What's more, some participants mentioned haptic feedback could distract DHH users like a notification system

instead of functioning as an aid to understanding the media content better. After distraction, switching focus back to haptic feedback and recalling speakers might increase the workload on users' memory. P6 commented that *" Distracting (is) like, I was trying to focus on the caption content, but then I feel the buzz and shift my attention to my watch and realized I forgot it mean switching speakers. So it took me times to focus on it and then try to remember which buzzing I'm feeling for which speakers."*

Using Haptic-Captioning in combination with other captioning methods: Many participants mentioned that the Haptic-Captioning system would be more beneficial by making it compatible with the non-real-time visual captions like position captions. P2 mentioned that: *"Haptic-Captioning adds another level of feeling connected to the content being shown, but I believe pairing that with positioned captions would be a great fit. Overall, I can see [Haptic-Captioning] being a thing if it's developed to be compatible with a user's preferred captioning method."* Similar to this point, P12 commented that the haptic feedback helps users better focus and enjoy the visual aspect of the media rather than paying attention to reading the text: *"Haptic-Captioning allows me to not focus on text so much in video, and appreciate the visual aspect more."*

Using Haptic-Captioning for other purposes than speaker indication: Some participants (P2 & P7) also expressed their desire to experience a using the Haptic-Captioning system with movies, music, and other genres. P7 mentioned: *"I wanted to wear the haptic device when I have a plan to watch a movie..to understand an ambiguous caption with the help of the haptic device."*

In addition to movies and music, P2 suggested using Haptic-Captioning for indicating other aspects of speech, such as emotions when watching visual content. Here, P2 mentioned, *"I definitely believe that adding haptic systems to music, movies and sports would be helpful as emotion is heard in people's voices. For example, having the ability to feel the intensity of how someone is speaking while a home run occurs in baseball (sports in general), or when someone is yelling in a movie, would be beneficial to deaf and hard of hearing people."*

5.6 Summary of Study 1

In this study, we compared the Haptic-Captioning system with several captioning methods, namely real-time and non-real-time captioning methods. Overall, we found that the Haptic-Captioning system performed best in identifying speaker transitions (not statistically significant). In addition, the Haptic-Captioning system was rated higher than the Traditional Real-time Caption in the post-trial questionnaire. However, the Position Caption has significantly better ratings when compared with the Haptic-Captioning, especially in terms of Confidence and Comfortableness dimensions. Here, we assume there is a significant learning effect that exists between visual and haptic cues while comparing the Haptic-Captioning system with non-real-time captioning methods. Some concerns from participants might also explain why the Haptic-Captioning system received a low rating compared to some visual captioning methods such as wearing tiredness and distraction caused by misunderstanding haptic feedback as a notification system. Moreover, the participant feedback indicated that the Haptic-Captioning has a great potential to complement the other limitations of non-real-time captioning methods, such as the distraction from moving text, inaccessibility for the color blind, etc. Participants also suggested using the Haptic-Captioning system in combination with other visual captioning methods. In addition, one main suggestion was to explore the system with different types of content and applications which is explored next in Study 2.

6 STUDY 2: CONTEXTUAL INTERVIEW

Inspired by the feedback from the last two studies, we aim to investigate the user experience of the Haptic-Captioning system in different contexts to access media sources of multiple speakers. Thus, we conducted a semi-structured contextual interview to elicit DHH users' feedback in different application settings and explore other factors that could inform the future design of the Haptic-Captioning system (RQ3).

6.1 Study Design

This study aims to understand DHH participants' experience with using the haptic modality in three different settings: TV, mobile, and laptop. Informed by Study 1, we made three video clips where each clip consisted of four genres of videos (i.e., podcast, sports, live stream, movie). Each video is approximately 4 minutes (1 min * 4 genres). We counterbalanced three settings to ensure each video played on different platforms. The video clips were played in the same order so that the disorder does not affect participants' understanding of the media contents. Before playing the video in each setting, we ask the participants to choose their preferred vibrating positions in Figure 1 A-D. This step was informed by participants' feedback from the first two studies that they would like to try out different positions instead of keeping it fixed on the wrist all the time. We gave three suggestions based on previous audio-haptic design wearing on the wrist [19], holding against the phone [24], and putting on the chair [38]. We designed a semi-structured interview to understand the user experience in these media settings and then explore the possibility of the future Haptic-Captioning system designs in terms of the context of use. The interview questions were focused on (1) Overall experience using the Haptic-Captioning system in terms of three settings, four video genres, and vibrating positions, (2) The benefits and challenges of haptic feedback in these settings, (3) The environmental factors and other different contexts of use that affect the experience (4) Suggestions for improving the design of the Haptic-Captioning system.

6.2 Participants

We recruited six participants, R1-R6 (three male, two female, one non-binary) aged 18-26 (M = 22, SD = 3.4) from Study 1. Participants' information (R1-R6) were shown in Table 1. Four participants reported profound hearing loss, one with mild hearing loss, and one with moderate hearing loss. As for the hearing devices used, three participants used hearing aid(s), two used Cochlear implant(s), and one did not use any hearing devices. Two participants preferred to communicate through speaking, and four participants chose to communicate through typing. According to the demographic questionnaire, all participants reported they had used captions on TV, mobile phone, and laptop. In terms of the familiarity of three

settings, all participants reported being extremely familiar with mobile phone and laptop settings. When it came to the TV setting, three participants were extremely familiar with it (R1, R4, R5, R6), one reported moderately familiar (R3), and one was slightly familiar (R2).

6.3 Study Procedure

In Study 2, each participant experienced using the Haptic-Captioning system on three devices with a four-minute video. The video clip contained four media genre types. Prior to the study, participants were asked to select the vibrating position they felt comfortable with using (see in Figure 1). We provided three positions as suggestions based on the comments from the previous studies on audio-haptic methods [19, 24, 38]. However, participants were free to change the positions to the ones they felt more comfortable with. Lastly, we conducted a semi-structured interview to understand the overall experience of using the Haptic-Captioning system regarding video genres, device settings, and vibrating positions. We asked questions related to their perceived benefits and challenges encountered and how the environment affects their experience. In the end, participants were encouraged to provide their suggestions to improve the system device. The contextual interview took approximately 50 minutes for each participant.

6.4 Results and Discussion of Study 2

We performed the thematic analysis with an open and inductive coding approach on the collected feedback [6]. One researcher scanned the raw transcripts and identified 133 comments from 6 participants (in a total of 6357 words). Then, one researcher developed initial open codes and shared them with the entire research team. We collaboratively generated the final open codes and then grouped them into themes. We used affinity diagrams on Miro[4] for searching emerged themes. We identified three themes which are *attention, perception, and customization*. We will present our major findings using the inductive themes and representative quotes below.

6.4.1 Attention. **Improve caption readability.** Through the post-study interview, participants explained further how haptic feedback benefited real-time captions and assisted readability. One characteristic of the Haptic-Captioning system was to foster ambient awareness of the media content. For example, R3 mentioned that haptic feedback helped the captions as a supplement by matching the textual input of words with the movement from lip reading. Thus, especially when people could not grasp information from lipreading like in podcasts, Haptic-Captioning would be more beneficial for ease of following:

> *"It helps me kind of identify the change of voice and kind of keep track of where I am with the captions. Like, I don't know if it was this unconscious thing, but I could kind of match up the vibration to the captions. When I was reading the captions, I could feel it as I was reading so I could tell where in the captions they were, kind of like, like some Disney lyric videos like the karaoke. You can follow along and there's a little bouncing, if you can tell which word they're on."*

[4]https://miro.com

Maintain media engagement. On the contrary, the Haptic-Captioning system could save user's attention while still helping them engage in media content. Taking the movie as an example, where participants sometimes did not focus on speaking much, they felt positive about the usage of the system, as R2 mentioned:

> *"I guess, say that you can understand the awareness of the movie. So I know like what's going on, [but] not necessarily what they're talking about. But seeing the action in the background, this would be helpful there. Yeah, but the speaking part, I just watch it like a typical movie."*

Similarly, R3 mentioned that she sometimes likes to remove the hearing aids so that she could focus less on the content. In this situation, the Haptic-Captioning system would help her engage more without requiring great attention:

> *"On occasion, when I'm really tired, I'll take out my hearing aids and like, watch Criminal Minds. And having this would help me, [to] get that input like the background noise, the sound. So whenever I take out my hearing aids, because I'm super tired, but I still want to be able to be engaged in the movie or the TV show, I would definitely go to that."*

6.4.2 Perception. Through the observation and the analysis of participants' comments, we identified that the Haptic-Captioning system could contribute to enhancing the perception of the vocal content for the users. Participants indicated that, with the assistance of the Haptic-Captioning system, they were able to perceive non-speech information, such as footsteps or raindrops, via various haptic patterns. Thus, they could be more engaged with the content and feel more confident in identifying the active speaker. In other words, the haptic system augmented their experience of watching vocal content by bringing more senses, especially in the context of watching movies. Here, participants noted, *"I found it to be a nice additional dimension to the media"* (R1); *"I think the movie will be a good experience using the Haptic-Captioning system because it provides better senses."* (R6)

Enhance understanding emotions. We found that the Haptic-Captioning system would benefit DHH people by enhancing their understanding of emotion based on different sound effects. For example, several participants mentioned they could feel the excitement as well as the scary sound effects and the laughing from ominous music. Specifically, while watching the movie, participants reported the Haptic-Captioning system was helpful for matching the actions that happened to the sound effects, which were hard for them to access. This observation was an indication that the Haptic-Captioning system could provide feedback on non-speech information. For example, R3 reported:

> *"It definitely helps make the emotions more easy to tell, because like, it's just kind of slowly vibrating. And then when they open the door and started running, it's like vibrating faster with the music and that kind of helped to match the motion at the scene to kind of the music that you otherwise wouldn't be able to hear, like the background sounds he wouldn't be able to hear."*

In the study, we found the majority of the participants preferred wearing the Haptic-Captioning system on the wrist (Figure 1-A). This observation aligned with the result from a previous survey on DHH users' preferences on wearable devices [12]. However, we also identified that some participants preferred using the Haptic-Captioning system by attaching it to the phone's back (see in Figure 1-C). The participants reported that this holding position allowed them to sense the vibration through their fingers. In addition, one participant elaborated on the differences in the experience of attaching the device to the phone and wearing it on the wrist. The participant stated that holding the phone with the haptic device in both hands allowed the user to perceive the vibrations more clearer. In this case, this holding position might be more preferable for users who wanted to be more engaged with the video content. It was supported by R04's comment:

> *"I think I am comfortable with haptic on the phone more than on the watch. I feel I can connect along with the phone and haptic at the same time compare to wrist. For some reason if the haptic system on my wrist, it feels disconnected somehow."*

Assist speaker indication. While watching the video without seeing the picture of speakers, such as in a podcast, participants mentioned that haptic feedback could help them identify the active speaker depending on the haptic patterns translating from the speaker's voice. As participants perceived a female voice felt softer while the male voice seemed deeper in general, it aligned with the participant's feedback from the Preliminary Study. However, participants also mentioned it could be a challenge to pick up the speaker's voice when the sound quality was low. Regarding this, R3 noted:

> *"I mean, in general, the live stream, it depended on the quality of the person's microphone that is speaking. So like the man who was doing the actual questioning and like the presenter guy, the news person, it was really clear to be able to pick out his voice. [The others], their microphones weren't that good. So there's just kind of a lot of constant vibration. And it wasn't succinct. It wasn't obvious."*

6.4.3 Customization. Although this experiment was carried out in a lab setting, we noticed that participants mentioned several social and environmental factors that could cause differences on their experience, such as social interaction, multitasking, distraction from external sound, etc. Here, we identified the customization needs of the Haptic-Captioning system to provide a better fit for the contexts mentioned above.

Support multitasking. Few participants mentioned their sensations would be different than watching TV with a group of people, as the scenarios might involve a variety of tasks. For example, haptic feedback might cause distraction in their conversation with others, while in the meantime, benefit from picking up TV content, as R1 commented:

> *"I think that since I'm sitting in a lab room my senses are very isolated. I would be curious to use this in a busy room of friends watching TV, and observe if I felt it was distracting me from conversation, or helping me to know when to turn back to the TV."*

Similarly, many participants suggested testing the Haptic-Captioning system built-in to a chair, which might specifically benefit the TV settings from freeing their hands. Based on our observation, we noticed that R1, who chose to place the Haptic-Captioning system on the leg (shown in Figure1-B), explained that although he preferred integrating the Haptic-Captioning system in the chair to not occupy his hands, the vibration of the device would be stronger when placing on the leg:

> *"For the TV it would make the most sense to have it built into a chair. That way it would still be able to provide strong vibrations despite external events, & you would not have to hold anything... in case you are signing with friends or something."*

Keep privacy in public. In some cases, participants would like to keep the Haptic-Captioning system more private from others. The apprehension of showing the Haptic-Captioning system in public would also affect the user preference on the wearing positions. For example, R3 mentioned she would be more cautious about using the device in the public environment:

> *"I think if I were in a public space, I would be more willing to kind of hold it against my phone. So it's less obvious. If I were in a private space, I would be more willing to move it around, and like test places sitting next to me on my wrist, like, see if there's a place that works best because I'm by myself or I'm in a private environment where people know that I use this."*

Avoiding environmental noise. R2 further elaborated on the difference between hearing in quiet and loud environments as the external noise would limit the hearing aid's abilities. For quiet settings, hard-of-hearing users might rely on their hearing more than on reading the caption. However, R2 later mentioned his experience as hard-of-hearing people would be varied from the deaf population:

> *"I think in a loud setting this is definitely helpful, because now they're not relying too much on hearing, rely more on text and such, and this one might actually be helpful. I think that's effect of environment. In a quiet environment, not necessary...I just keep hearing this. But in a loud settings, probably, probably better."*

However, the vibration from the environment might cause confusion on the understanding of the haptic information. R4 explained the distraction that prevented them from understanding the information conveyed by the Haptic-Captioning system in a public transportation:

> *"Suppose if I was in a car or subway, it may affect my experience with the haptic feedback while watching the stream. Transportation tend to have vibration such as loud engine or bump that cause the vibration or movement. It may conflict with my experience while watching. Let's say if I'm holding my phone with the haptic system while in the subway (Without hearing aids), I can get confused if the subway has an announcement while watching the video with haptic feedback."*

6.5 Summary of Study 2

Study 2 is designed to elicit formative feedback from participants on the Haptic-Captioning system and explore the situations of this multimodal system that would increase caption accessibility. Participants reflected on several positive experiences of using the Haptic-Captioning system as it improved the caption's readability and maintained media engagement. The haptic feedback led them to unconsciously engage in the content whether or not they needed to understand the text. Moreover, the Haptic-Captioning combined with the traditional real-time system enhanced the user perception in the aspect of emotions and then assisted speaker indication, especially when the speaker in the media was invisible. On the other side, we also identified several situations that the system design needs to be improved due to social and environmental factors. As participants discussed their concerns, we discover there are design opportunities to customize the Haptic-Captioning system to fit different contexts. Some expressed their preferences on wearing positions and usage environments might change according to the context of use, which required future examination. We will discuss the design implication in detail in Section 7.1.

7 DISCUSSION

In this paper, we first proposed Haptic-Captioning system (Figure 1) and investigated how Haptic-Captioning system assisted DHH users with speaker indication in multiple-speaker media. Our three user studies illustrated the potential of using haptics to convey the speaker's information that aimed to improve the accessibility and understanding of captioning. Below, we present our takeaways reflected from our findings and then discuss the implication for the Haptic-Captioning system design.

7.1 Takeaways from the Studies and improvements for the Haptic-Captioning system

Our takeaways demonstrate several factors related to the efficacy of using the Haptic-Captioning system on speaker indication, which should be considered in future design.

Similarity between speakers' voice patterns. Our findings from our three studies suggested that DHH participants found the difficulties of identifying speakers varied from the number of speakers and their background, which extended the challenge identified in previous work [36]. In the Preliminary Study which we only examined the haptic feedback, participants were able to identify the total number of speakers with over 70% mean accuracy. However, the mean accuracy in the trials of two speakers is significantly different from with three speakers. Similar feedback was also observed in Study 1 and Study 2 with reference to background sounds, overlapping conversations, etc.

Familiarity towards haptic and media. Participants' feedback in Study 1 revealed that the level of experience affected their preferences in general. The learning curve to familiarizing with the haptic pattern might bring challenges to DHH users that lowered their confidence in speaker indication. However, we believe such challenges could be tackled through a longer exposure to the haptic modalities. Similarly, in Study 2, one participant explained that the level of familiarity with the haptic feedback benefits his understanding of testing this device. R4 commented that his hearing aids enable him to be acquainted with the sound pattern of the environment. The familiarity factor is not just identified in the haptic patterns, but also in the channel for accessing the media. Here, few participants mentioned their relatively low frequency of watching TV compared to using a phone or a laptop. Therefore, our next step in this direction is to provide haptic training or design a longitude study to explore the Haptic-Captioning in depth.

Attention required on the visual information. Our study extended on existing literature on the visual captioning style preference in terms of comparing with the Haptic-Captioning modality [1, 3, 4]. From the quantitative data, we did not find any statistically significant difference in speaker transition's mean accuracy between haptic captions and non-real-time captions. The comparison of haptic and non-real-time captions revealed that while DHH people generally prefer visual cues, these extra adds-on might raise new challenges on the increased eye fatigue and distraction, which could impact the readability of captions [21]. With the Haptic-Captioning system, participants in Study 1&2 mentioned they have a chance to enjoy the content itself rather than focusing on the captions. In some cases, when users prefer to play sounds as background noise, combining the Haptic-Captioning with an appropriate visual method helps maintain the awareness of the environment. Future studies should examine the combination of haptic and visual captioning and examine how Haptic-Captioning complements the visual aspect of media.

7.2 Future Design Implications of Haptic-Captioning System

Our findings indicate the future design of the wearable Haptic-Captioning system should be comfortable, understandable, and transportable. We identified three main future research directions based on the Haptic-Captioning system and its new uses.

Firstly, the Haptic-Captioning system could be improved in providing adjustable vibration. DHH participants switched positions several times during Study 2 to adjust the intensity of the vibration, specifically when watching movies and sports, which they desired to receive more feedback on the non-speech information. However, an intense vibration would also cause sound leakage, which some participants were worried about using the Haptic-Captioning system in the public scenario. Besides, some participants also commented that a strong vibration would bring fatigue for a longer time use. Therefore, participants should be able to customize the volume level to fit their needs. The customization should also help users distinguish the haptic pattern between notification reminders and media content aids. While many works have attempted integrating haptic devices in such contexts and attached to mobile[24], we are motivated to explore this in a captioning context.

Secondly, our participants suggested that the Haptic-Captioning system should be transportable like a wristband device. For example, a haptic wristband could build upon the wearable haptic device for the hands [33]. We aimed to explore a wristband prototype that provides spatial haptic feedback with multiple actuators where the haptic feedback is associated with the position of the speakers. It is also important to develop a sound-haptic algorithm that can standardize the audio in real-time with a separate sound channel. Some

participants suggested that future design could consider a build-in system to provide a more immersive experience without occupying their hands. For the built-in system, the future design could integrate the auditory-haptic vibrator in chairs, game controllers, and mobile phones, which could bring a full-body experience to DHH users.

Thirdly, inspired by the participant's feedback, we aimed to explore how the Haptic-Captioning can present feedback to convey the tones and emotions of speakers. During the studies, several participants briefly mentioned that they could potentially identify whether the speakers were speaking in an angry tone or sad tone. While presenting non-speech information had been explored in the past with haptic feedback [23], we posited that our method would be able to present such "meta" speech information to DHH users as well. Thus, this is a major research direction we aim to explore in this work.

7.3 Limitation & Future work

While the participants' demographic such as level of hearing ability is always interesting to investigate [16], in this study, we tried to tackle this factor by putting headphones with the white noise in the Preliminary Study and turning off the external sound in Study 1. During Study 2, one hard-of-hearing participant mentioned the variation of hearing ability might affect their perception in the public scenario. In future work, the demographic and prior experience of DHH participants should be considered as factors of their preferences.

In the preliminary study, our goal is to understand DHH viewers' perception of speaker information with haptic feedback. We intend to investigate the user perception of general demographic information about the speaker. However, the majority of participants tended to answer this question related to gender. We observe participants have thoughts on how haptic patterns related to the binary gender (i.e., males are more likely to have the low pitch). We assume haptic modality has the potential to avoid prejudging the gender of the speaker from the real-time captioning method. More future will be needed to examine the DHH viewers' understanding of broader gender information through the multi-modal presentation.

Regarding the ecological validity of Study 2, we put efforts into setting up the contextual interview in a comfortable lab environment with three settings (i.e., TV, laptop, phone) and four genres (i.e., podcast, sports, live stream, movie). This arrangement was able to help us elicit users' feedback on the Haptic-Captioning system while considering different multiple-speaker scenarios. Still, we acknowledge that the isolated lab could constrain participants' holistic media experience. More work is needed to examine the DHH users' experience using the Haptic-Captioning system to access multi-speakers media sources in real-world contexts. Besides, we do not claim our finding is applicable to support speaker identification in real-time communication such as small group conversation [29]. This is another direction we wish to explore in the future to use the Haptic-Captioning system in in-person and virtual meetings with multiple speakers. Some social and environmental factors with the Haptic-Captioning should be considered, ranging from users' communication styles and the environment's dynamic to noise-canceling features of video conferencing tools.

8 CONCLUSION

This study has investigated Haptic-Captioning system through a three-phase experiment. Our preliminary study revealed interesting insights on the user perception of the speaker information with only haptic feedback. Next, through a within-subjects study with 16 DHH participants, we compared the Haptic-Captioning system with the existing visual modalities. We found that there was no significant difference in identifying speaker transition between haptic and visual captioning methods. In terms of subjective experience, the Haptic-Captioning system has more positive ratings than the traditional real-time caption's ones among all dimensions, but no significant difference was founded between those two conditions. Lastly, we conducted a contextual interview to understand user experience using the Haptic-Captioning system in three semi-realistic settings (TV, mobile, laptop) and observed participants' preferences on wearing positions. Our qualitative data analysis suggested the overall characteristics of the Haptic-Captioning system and informed the future direction of design and research.

ACKNOWLEDGMENTS

This research was supported by the seed funding award from the iSchools Inc. Caluã de Lacerda Pataca is considered as the 4th author of the paper.

REFERENCES

[1] Akhter Al Amin, Abraham Glasser, Raja Kushalnagar, Christian Vogler, and Matt Huenerfauth. 2021. Preferences of Deaf or Hard of Hearing Users for Live-TV Caption Appearance. In *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments*, Margherita Antona and Constantine Stephanidis (Eds.). Springer International Publishing, Cham, 189–201.

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

[3] Akhter Al Amin, Joseph Mendis, Raja Kushalnagar, Christian Vogler, Sooyeon Lee, and Matt Huenerfauth. 2022. Deaf and Hard of Hearing Viewers' Preference for Speaker Identifier Type in Live TV Programming. In *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments*, Margherita Antona and Constantine Stephanidis (Eds.). Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-05028-2_13

[4] Larwan Berke, Khaled Albusays, Matthew Seita, and Matt Huenerfauth. 2019. Preferred appearance of captions generated by automatic speech recognition for deaf and hard-of-hearing viewers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[5] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. 2017. Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One Meetings. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 155–164. https://doi.org/10.1145/3132525.3132541

[6] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. (2012).

[7] Andy Brown, Rhia Jones, Michael Crabb, James Sandford, Matthew Brooks, Michael Armstrong, and Caroline Jay. 2015. Dynamic Subtitles: The User Experience. https://doi.org/10.1145/2745197.2745204

[8] Janine Butler. 2020. The Visual Experience of Accessing Captioned Television and Digital Videos. *Television & New Media* 21, 7 (2020), 679–696. https://doi.org/10.1177/1527476418824805 arXiv:https://doi.org/10.1177/1527476418824805

[9] Angela Chang and Conor O'Sullivan. 2005. Audio-haptic feedback in mobile phones. In *CHI'05 extended abstracts on Human factors in computing systems*. 1264–1267.

[10] Artem Dementyev, Pascal Getreuer, Dimitri Kanevsky, Malcolm Slaney, and Richard F Lyon. 2021. VHP: Vibrotactile Haptics Platform for On-Body Applications *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 598–612. https://doi.org/10.1145/3472749.3474772

[11] Described and Captioned Media Program. 2022. Captioning key - speaker identification. https://dcmp.org/learn/603-captioning-key---speaker-identification. Accessed: 2022-04-12.

[12] 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

[13] Morton Ann Gernsbacher. 2015. Video Captions Benefit Everyone. *Policy Insights from the Behavioral and Brain Sciences* 2, 1 (2015), 195–202. https://doi.org/10.1177/2372732215602130 arXiv:https://doi.org/10.1177/2372732215602130 PMID: 28066803.

[14] Abraham Glasser, Edward Mason Riley, Kaitlyn Weeks, and Raja Kushalnagar. 2019. Mixed Reality Speaker Identification as an Accessibility Tool for Deaf and Hard of Hearing Users. In *25th ACM Symposium on Virtual Reality Software and Technology* (Parramatta, NSW, Australia) *(VRST '19)*. Association for Computing Machinery, New York, NY, USA, Article 80, 3 pages. https://doi.org/10.1145/3359996.3364720

[15] 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*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376406

[16] Stephen R Gulliver and George Ghinea. 2003. How level and type of deafness affect user perception of multimedia video clips. *Universal Access in the Information Society* 2, 4 (2003), 374–386.

[17] Richang Hong, Meng Wang, Xiao-Tong Yuan, Mengdi Xu, Jianguo Jiang, Shuicheng Yan, and Tat-Seng Chua. 2011. Video Accessibility Enhancement for Hearing-Impaired Users. *ACM Trans. Multimedia Comput. Commun. Appl.* 7S, 1, Article 24 (nov 2011), 19 pages. https://doi.org/10.1145/2037676.2037681

[18] Yongtao Hu, Jan Kautz, Yizhou Yu, and Wenping Wang. 2015. Speaker-Following Video Subtitles. *ACM Trans. Multimedia Comput. Commun. Appl.* 11, 2, Article 32 (jan 2015), 17 pages. https://doi.org/10.1145/2632111

[19] 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 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

[20] Dhruv Jain, Rachel Franz, Leah Findlater, Jackson Cannon, Raja Kushalnagar, and Jon Froehlich. 2018. Towards Accessible Conversations in a Mobile Context for People who are Deaf and Hard of Hearing. 81–92. https://doi.org/10.1145/3234695.3236362

[21] Kuno Kurzhals, Emine Cetinkaya, Yongtao Hu, Wenping Wang, and Daniel Weiskopf. 2017. Close to the Action: Eye-Tracking Evaluation of Speaker-Following Subtitles. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6559–6568. https://doi.org/10.1145/3025453.3025772

[22] Raja Kushalnagar, Gary Behm, Kevin Wolfe, Peter Yeung, Becca Dingman, Shareef Ali, Abraham Glasser, and Claire Ryan. 2019. RTTD-ID: Tracked captions with multiple speakers for deaf students. *arXiv preprint arXiv:1909.08172* (2019).

[23] Raja S. Kushalnagar, Gary W. Behm, Joseph S. Stanislow, and Vasu Gupta. 2014. Enhancing Caption Accessibility through Simultaneous Multimodal Information: Visual-Tactile Captions *(ASSETS '14)*. Association for Computing Machinery, New York, NY, USA, 185–192. https://doi.org/10.1145/2661334.2661381

[24] Jaebong Lee and Seungmoon Choi. 2013. Real-time perception-level translation from audio signals to vibrotactile effects. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2567–2576.

[25] Tomosuke Maeda, Roshan Peiris, Nakatani Masashi, Yoshihiro Tanaka, and Kouta Minamizawa. 2016. HapticAid: Wearable Haptic Augmentation System for Enhanced, Enchanted and Empathised Haptic Experiences. In *SIGGRAPH ASIA 2016 Emerging Technologies* (Macau) *(SA '16)*. Association for Computing Machinery, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/2988240.2988253

[26] Tomosuke Maeda, Roshan Peiris, Masashi Nakatani, Yoshihiro Tanaka, and Kouta Minamizawa. 2016. Wearable Haptic Augmentation System Using Skin Vibration Sensor. In *Proceedings of the 2016 Virtual Reality International Conference* (Laval, France) *(VRIC '16)*. Association for Computing Machinery, New York, NY, USA, Article 25, 4 pages. https://doi.org/10.1145/2927929.2927946

[27] Tomosuke Maeda, Keitaro Tsuchiya, Roshan Peiris, Yoshihiro Tanaka, and Kouta Minamizawa. 2017. Hapticaid: Haptic experiences system using mobile platform. In *Proceedings of the Eleventh International Conference on Tangible, Embedded, and Embodied Interaction*. 397–402.

[28] Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications* 80, 6 (01 Mar 2021), 9411–9457. https://doi.org/10.1007/s11042-020-10073-7

[29] Emma J. McDonnell, Ping Liu, Steven M. Goodman, Raja Kushalnagar, Jon E. Froehlich, and Leah Findlater. 2021. Social, Environmental, and Technical: Factors at Play in the Current Use and Future Design of Small-Group Captioning. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 434 (oct 2021), 25 pages. https://doi.org/10.1145/3479578

[30] Kouta Minamizawa, Yasuaki Kakehi, Masashi Nakatani, Soichiro Mihara, and Susumu Tachi. 2012. TECHTILE toolkit. In *IEEE Haptics Symposium*.

[31] Suranga Nanayakkara, Elizabeth Taylor, Lonce Wyse, and S H Ong. 2009. An enhanced musical experience for the deaf: design and evaluation of a music display and a haptic chair. In *Proceedings of the sigchi conference on human factors in computing systems*. 337–346.

[32] Andrew J. Oxenham. 2018. How We Hear: The Perception and Neural Coding of Sound. *Annual Review of Psychology* 69, 1 (2018), 27–50. https://doi.org/10.1146/annurev-psych-122216-011635 arXiv:https://doi.org/10.1146/annurev-psych-122216-011635 PMID: 29035691.

[33] Claudio Pacchierotti, Stephen Sinclair, Massimiliano Solazzi, Antonio Frisoli, Vincent Hayward, and Domenico Prattichizzo. 2017. Wearable haptic systems for the fingertip and the hand: taxonomy, review, and perspectives. *IEEE transactions on haptics* 10, 4 (2017), 580–600.

[34] Frank A Saunders, William A Hill, and Barbara Franklin. 1981. A wearable tactile sensory aid for profoundly deaf children. *Journal of Medical Systems* 5, 4 (1981), 265–270.

[35] Radu-Daniel Vatavu. 2021. Accessibility of Interactive Television and Media Experiences: Users with Disabilities Have Been Little Voiced at IMX and TVX. In *ACM International Conference on Interactive Media Experiences* (Virtual Event, USA) *(IMX '21)*. Association for Computing Machinery, New York, NY, USA, 218–222. https://doi.org/10.1145/3452918.3465485

[36] Quoc V Vy and Deborah I Fels. 2010. Using placement and name for speaker identification in captioning. In *International Conference on Computers for Handicapped Persons*. Springer, 247–254.

[37] Maximilian Weber and Charalampos Saitis. 2020. Towards a framework for ubiquitous audio-tactile design. In *International Workshop on Haptic and Audio Interaction Design*. Montreal, Canada. https://hal.archives-ouvertes.fr/hal-02901209

[38] Antoine Weill–Duflos, Feras Al Taha, Pascal E. Fortin, and Jeremy R. Cooperstock. 2019. BarryWhaptics: Towards Countering Social Biases Using Real-Time Haptic Enhancement of Voice. In *2019 IEEE World Haptics Conference (WHC)*. 365–370. https://doi.org/10.1109/WHC.2019.8816153

[39] J M Weisenberger, S M Broadstone, and L Kozma-Spytek. 1991. Relative performance of single-channel and multichannel tactile aids for speech perception. *J Rehabil Res Dev* 28, 2 (1991), 45–56.

[40] J M Weisenberger, S M Broadstone, and F A Saunders. 1989. Evaluation of two multichannel tactile aids for the hearing impaired. *J Acoust Soc Am* 86, 5 (Nov. 1989), 1764–1775.