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

by

Yiwen Wang

Project submitted in partial fulfillment of the requirements for the

degree of Master of Science in Human-Computer Interaction

Rochester Institute of Technology

B. Thomas Golisano College
of
Computing and Information Sciences

Department of Information Sciences and Technologies

Rochester Institute of Technology

B. Thomas Golisano College
of
Computing and Information Sciences

04/26/2022

Master of Science in Human-Computer Interaction

Rochester Institute of Technology

B. Thomas Golisano College

of

Computing and Information Sciences

Master of Science in Human-Computer Interaction

~ Project Proposal Approval Form ~

Student Name: Yiwen Wang

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

Project Area(s):	☐ Application Dev.	☐ Database	☐ Website Dev.
(√ primary area)	☐ Game Design	x HCI	☐ eLearning
	☐ Networking	☐ Project Mngt.	☐ Software Dev.
	☐ Multimedia	☐ System Admin.	☐ Informatics
	☐ Geospatial	☐ Other _____	

~ MS Project Committee ~

Name	Signature	Date
Chair: Dr. Roshan Peiris		
Committee: Dr. Tae Oh		

Table of Contents

1 Abstract

Captions are developed to make the audio content of videos accessible and understandable for people who are deaf and hard-of-hearing (DHH). While text-based captioning methods are often used, traditional captioning remains challenging for DHH users who have difficulty distinguishing the characteristics of sound to identify the active speaker in multiple-speaker scenarios. In order to enhance the accessibility of captioning, we proposed a Haptic Captioning system that provides real-time vibration feedback on the wrist by directly translating the sound output. In this paper, we conduct three-phase experiments to examine haptic perception (Preliminary Study), compare the haptic modality with visual-based captioning methods (study 1), and then investigate the user experience of using the Haptic Captioning system through a contextual interview (Study 2). Although Study 1 suggests that DHH users prefer visual caption modalities, Study 2 further found that the Haptic Captioning is able to complement the visual captions by enhancing emotion understanding, improving caption readability, and assisting speaker indication, especially in real-time captioning scenarios.

2 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 an environment such as a live discussion with multiple panelists or viewing 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. 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, 21].

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 [32], adding the avatar image of speakers [32], highlighting the active speaker through pop-ups [21], and signifiers [14]). However, these 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, 25]. In addition, concerns have been discussed about the extra cognitive loads required with visual speaker's 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 [21].

To address the challenges above, we propose Haptic Captioning System, an audio-haptic based real-time captioning system (Figure 1). The Haptic Captioning system directly translates the auditory content into haptic vibrations using a voice-coil a haptic actuator that is powered by an audio power-amplifier [26]. As such, this method generates vibrations that carry the same properties as the audio signal, preserving characteristics such as the frequency and timbre of the voices. Therefore, 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 audio-haptic systems have been frequently utilized to provide sound awareness for DHH users based on the characteristics of the sound [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 [22].

In this research, we explored the Haptic Captioning system as a wearable, holdable or attachable device for providing enhanced haptic feedback for captions (Figure 6). 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 of the Haptic Captioning system with 12 DHH participants. The participants were asked to wear the Haptic Captioning device on their wrists and discuss the identified speaker demographic by listening to audio clips via vibrations. Participants also reported their understanding of the speaker information, such as the number of speakers and their demographics. 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 six types of visual speaker indication methods. Study 2 focused on a qualitative approach

with a contextual interview by providing users with different genres of content (podcast, sports, live stream, movie) and settings (i.e., mobile, TV, laptop) together with the Haptic Captioning system.

In summary, our main contributions include:

- The Haptic Captioning system that directly translates audio information into haptic patterns to enhance captions.
- A preliminary investigation of the Haptic Captioning system's speaker indication characteristics with DHH participants.
- A comparison and discussion on speaker indication accuracies and the user preferences on haptic and visual captioning modalities.
- A qualitative finding on DHH people's experience of Haptic Captioning in different contexts of use and design implication for future Haptic Captioning device.

3 Prior Work

This section discusses the previous research done on caption accessibility, speaker indication methods in a captioned video, and techniques used for auditory-haptic translation.

Speaker Indication Accessibility and Challenges

Over the years, previous work stressed the importance of video accessibility through captions which proves to improve attention and comprehension for people learning to read, understanding the non-native languages, and for DHH people [13]. To enhance the accessibility of video content for DHH people, a prior study [8] explored how captions can influence the viewing experience. Qualitative analysis of the study discusses the viewing balance of caption and action, 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 caption based on individual preference and their engagement with the visual-aural content. Another research on improving caption accessibility examines 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 the caption accessibility where a study by Quoc et al., underlined the difficulty in identifying speaker change for media content that has multiple speakers, narrative discussions, and off-screen speakers [32].

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, 21]. 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]. Glasser et al., 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]. 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 [32]. 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 speakers' name in a video.

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 [28]. Audio-haptic technologies have been widely used in recent research for a wide range of applications [9, 10, 24, 26]. Among these, many works have explored using audio-haptic or sound-based haptic to make every day sounds accessible for DHH users [27, 31]. To enhance the sound awareness of DHH people, tactile technology has been used to identify the sound patterns through a wrist-worn device that emits haptic feedback based on the sound level [12, 19]. This study being one of the primary motivations behind our work suggest that vibrotactile information enhances the sound "experience" in the environment through evaluation in a life field experiment. Research has also been done on enhancing caption accessibility through visual-tactile information [22]. 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 [33] where haptic sensory was placed on various places such as armrest, back-rest [27], and under the seat [34].

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 several research questions we would like to answer in this study.

- RQ1: How can DHH viewers perceive speaker information through haptic feedback?
- RQ2: What are the user preferences and efficiency of speaker indication Haptic Captioning modality in comparison to existing visual captions?
- RQ3: How Haptic Captioning system affect DHH's user experience and what factors should be considered in future designs?

4 System Design

The Haptic Captioning system uses ***voice coil actuators*** to present the haptic vibration shown in Figure 1. Voice-coil actuators are vibrotactile devices that vibrate using sound signals. These devices are similar to an audio speaker but without the speaker's cone that amplifies the sound (therefore, voice-coil actuators may emit a slight sound when in use). In this research, we use the Acouve Vp210 actuator as our voice coil actuator. To actuate the actuator, we use a power amplifier based on the Techtile Toolkit design [26] (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 using any system. 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 that was looped through the casing to allow the participants to easily wear and take off the actuator similar to a wristwatch. In addition, the casing can be easily held or attached to a device such as a mobile phone-based on the requirement.

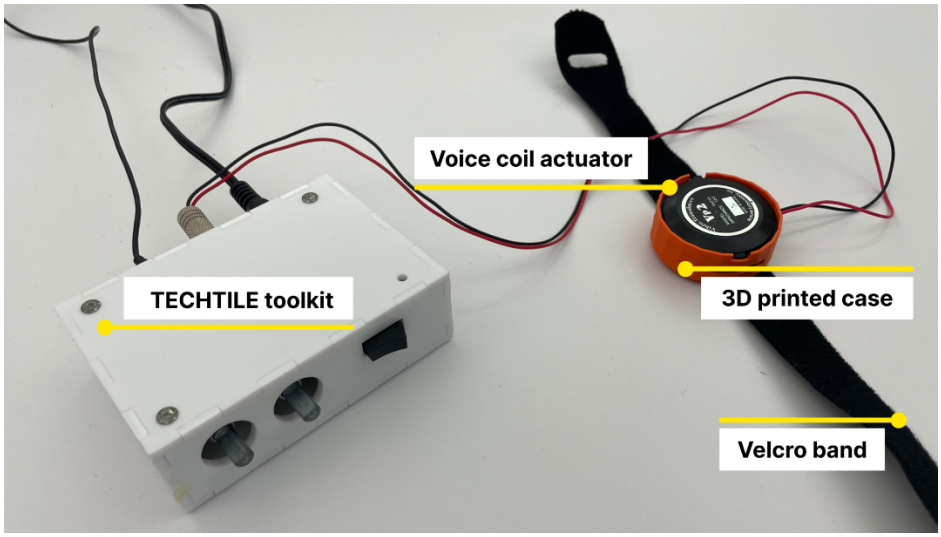

Figure 1. Haptic Captioning system consists of a voice-coil actuator, TECHTILE toolkit, 3D printed case, and Velcro band

5 Preliminary Study

The goal of this Preliminary Study was to gain initial insights into the characteristics of the Haptic Captioning system since our method proposes the use of haptic vibrations for speaker indication (RQ1). Therefore, we explored the DHH user's perception specifically on the haptic feedback generated from audio recordings of multiple-speaker panel discussions. All studies listed in this paper were approved by the Ethics Board of the Institution. In all studies, each participant was paid $15 for participation.

Study Design

To answer RQ1, we designed a within-subject study to examine the haptic perception. We selected 16 audio clips without any captions and visuals. We removed any captions and visuals as we wanted the participants to focus **only** on the haptic feedback and avoid being biased by the content of the captions or visuals (the video zooming on speakers and/or lip reading). Thus, we recorded sixteen 1-minute audio clips from eight live stream videos on YouTube, in which speakers discussed a range of topics including sustainability, life wisdom, business, education, and fashion. When selecting the clips, we included different levels of complexity based on the total number of speakers and the demographics. These complexities were decided based on trial and error experiments within the research team that also included one DHH member. The number of speakers ranged from 2-3 in the discussion (two speakers: 8 trials; three speakers: 8 trials). Demographics of speakers vary from age, perceived gender, ethnicity, etc. Inspired by previous work on transition cues [10], eight clips (four from each category) removed back-channel cues like "well", "em" and the other eight clips were recorded directly from YouTube. Participants were informed about a wider range of the total number of speakers and possible demographics before starting the study. The sequence of present haptic-auditory trials was randomized in the study software.

Apparatus

As our prototype, we investigated the wrist-band type Haptic Captioning prototype in the Preliminary Study (Figure 2). Several studies have demonstrated the haptic sensitivity of the skin on the wrist which motivated us to explore this site in our Preliminary Study [10]. In addition, the majority of the new smartwatches consist of a haptic vibration system that makes it easy to adapt our system as a future wearable device. The study software was programmed using C\# in Unity and was presented on a Macbook Pro 13" laptop.

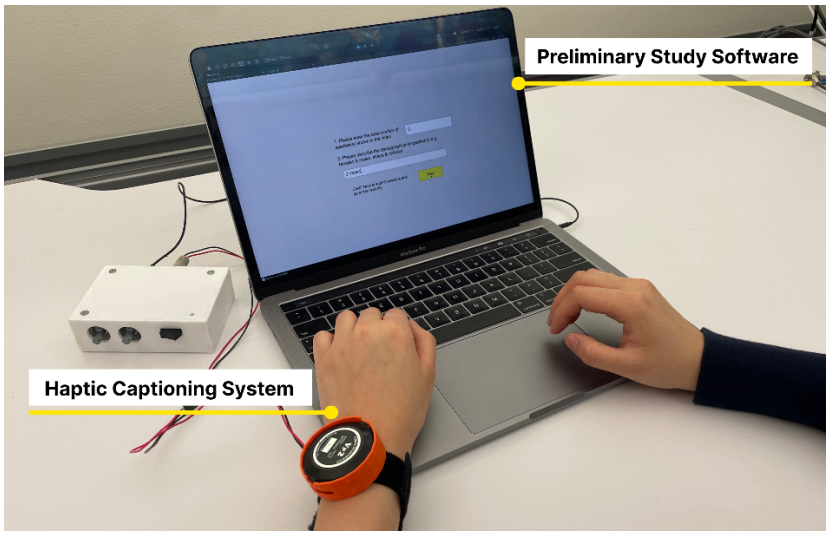

Figure 2. The Preliminary Study set-up consists of Haptic Captioning system that provides auditory-haptic feedback and the study software which was used to collect the perceived demographic information.

Participants

We recruited twelve DHH volunteers (five females, six males, and one non-binary) between the ages of 18 and 44 (M = 26.3, SD = 7.3) from social media, institute mailing lists as well as snowball sampling. Seven participants identified themselves as with profound hearing loss, and five with mild or severe hearing loss. Six participants reported they used hearing aid/s, one participant used Cochlear implant/s, and one participant used both. Four participants used none of the hearing devices and one participant used transcribers. For their capability of lip-reading, five participants reported being extremely familiar, and five participants were at least slightly familiar, two participants were not familiar at all. Each participant received a $15 reward for participation in this study.

Study Procedure

After introducing the system, the study procedure, and the experiment software, we set up the wearable vibrator on the participant's wrist. Inspired by the previous study [14], we asked the participants to put on a headphone that played white noise after gaining their consent. This was done to avoid any leaked sound from the voice coil actuator being heard by the participants. Before the trials began, we played different audio clips to demonstrate how the system worked. As the task began, the experiment software played a random audio clip with the haptic feedback. The participant could pause or play the clip or stop it at any time. As the task, after played the clip, the participant was taken to a page where the perceived total number of speakers and the demographic information were entered (Figure 2). We conducted a debriefing session to understand their overall experience, especially their haptic perceptual process of recognizing the switches and demographics of speakers. Each experiment took approximately 45 minutes.

Results and Discussion of Preliminary Study

The quantitative data in this study mainly include the total perceived number of speakers and the perceived demographic. We collected the feedback from DHH participants in the post-study session.

Speaker demographic

The accuracy of identifying the number of speakers was calculated by the perceived number of speakers divided by the actual number of speakers. The overall mean of the accuracy of the number of speaker predictions is 72.34%. We further computed the accuracy of the perceived number of speakers in two groups: 2 speakers (N=8) and three speakers (N=8). We found that 2 speakers (Mean=79.168%, SD =6.514) and 3 speakers (Mean=70.921%, SD =5.851). We further conducted a one-way ANOVA test and found there is a significant difference between the results of the group of 2 speakers and three speakers ($F(1,14) = 7.097$, $p<.05$). There was no significant difference for the overlapping ($F(1,14) = 2.616$, $p=.128$).

As for the perceived demographics of the speakers, the participants were free to report any characteristics they identified about the speakers. Hence, the participants reported various information such as their perceived gender and age (older adult, child, etc.) that were coded, analyzed and compared with the demographic information provided in the videos. We calculated the overall mean of the accuracy of the perceived demographic (i.e. gender) is 48.32%. More specifically, the two speakers' gender identification accuracy (N=8) is 52.77%, 3 speakers' gender identification accuracy (N=8) is 43.52%. We also found that the two clips in particular that participants were able to identify with over 70% accuracy both consisted of two speakers: one video with (one male older person and one male child) and another one with (one male adult and one female adult). Further, participants' feedback indicated how they identified the demographics of the speaker as observed in the comment below

---

*"Heavy buzz or high buzz makes me think adult and male... high light buzz makes me think of female. Low softer buzz makes me think of children. Sometimes that messes me up by having a combination of buzz makes me think of either softer buzz and high pitch could be female that speaker so loud... Not sure really. I just know the heavy buzz is male and the light buzz is female." – **P09***

One clip reported accuracy of less than 30%, which has three male adults. Supported by the feedback from post-study debriefing, participants mentioned the challenge of recognizing multiple speakers, especially when speakers have voice patterns like tones.

"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 - were they laughing? Were they just simply changing their tones? These are some questions that I thought of." – **P05***

Based on the above comment and similar comments from other participants, we identify the importance of the characteristics of the individual speaker's voice as well as the tone of the voice (emotion). For example, one speaker could escalate the tone and change the emotion, making the speaker identification difficult. However, as more channels such as visuals and captions are presented as per usual, we identify that speaker identification can be enhanced using our method.

Summary of the Preliminary Study

In this Preliminary Study, we examined how haptic vibrations alone could be used to convey demographic information about speakers. Overall, we observed that the participants identified demographics with relatively high accuracy (above 70 %) in terms of the number of speakers based on the vibration feedback alone. However, the accuracy decreased when the number of speakers increased. Furthermore, identifying other demographic information such as the perceived gender was found to be challenging as the voice characteristics are primarily dependent on the individuals. Thus, our next step is to explore the performance of the Haptic Captioning system at speaker indication with visuals and captions as a multimodal feedback system.

6 Study 1: Caption Methods Comparison

Our Preliminary Study indicated that audio-generated haptic feedback alone was promising at speaker indication. However, in an ideal situation, the Haptic Captioning system would be used in combination with the visuals of the content and their real-time captions. Thus, the main goal of this study is to explore the feasibility of the Haptic Captioning method in combination with visuals and captions. In addition, we compare it with prior visual caption methods and traditional captioning styles (RQ2).

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. Caption Modality consisted of seven conditions: the proposed Haptic Captioning method, six visual captioning styles shown in Figure 3 (i.e. avatar, color, position, pointer, speaker name, and traditional captions as the baseline condition). The Haptic Captioning condition used traditional captions style in combination with the system but used with a different video.

- ***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 [32].
- ***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].
- The ***Haptic Captioning*** provide the tactile feedback generated from the auditory content with the traditional plain text as shown in Figure 3 (f).

The visual content was 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 caption modality conditions were randomized. In total, each participant tested seven trials.

Perceived Speaker Transitions

Most previous visual caption methods used user ratings to evaluate user preferences and experiences of different captioning methods [21, 22, 32]. Therefore, to quantitatively evaluate the efficacy of the captioning methods, we explore 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. Thus, a participant may use the presented information from the different channels (visual, audio, haptic, etc.) to identify the speaker and thus 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.) where some caption methods

such as colored captions could, but, methods such as position captions could not indicate the speaker. 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 compare the time at which the participant reported a perceived speaker indication and sum all the correctly perceived speaker transitions events. The accuracy is defined as the percentage of correctly perceived speaker transition events over the actual number of events (from the coded data).

Figure 3: Visual Caption modalities: (a) avatar caption, (b) color caption, (c) position caption, (d) pointer caption, (e) speaker-name caption, (f) traditional caption, the same style but a different video used with Haptic Captioning.

Participants

We recruited sixteen participants (nine male, six female, one non-binary) aged 18-35 (M = 23.8, SD = 4.7) from social media and the institute's mailing lists. Ten participants reported they had profound hearing loss, three as mild hearing loss, two as severe, and one had moderate hearing loss. As for the hearing devices used, six participants used hearing aid/s, three used Cochlear implant/s, and one used both. Six participants used none of the hearing devices. For the level experience of lip-reading, five participants reported being extremely familiar, and eight participants were at least slightly familiar, three participants were not familiar at all.

Apparatus

For this study, we used the same wearable prototype that was used in the Preliminary Study (Figure 4). In the experiment software (developed using C# in Unity 3D), we presented the participant with the video with the selected captioned conditions. Here, we also added a button right next to the video to click whenever the participants identified a speaker transition (Figure 4) the procedure is discussed more in the following sections.

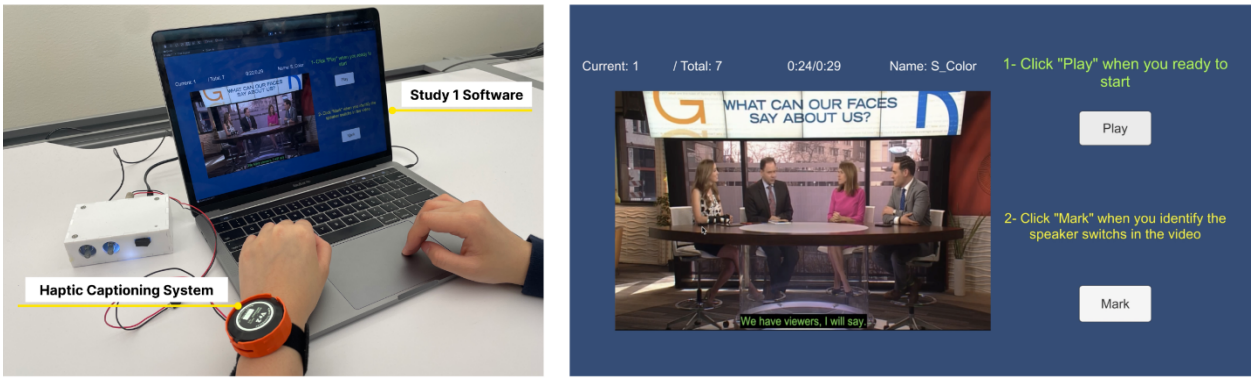

Figure 4: (Left) Study 1 set-up consists of Haptic Captioning system that only turns on in the Haptic Captioning condition and (Right) the screenshot of the study 1 software interface to log the "speaker transition" data

Study Procedure

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 is about 30 seconds. The introduction and demonstration allow participants to familiarize themselves with the task. Participants can 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 selected. Next, they used their other hand to click a "Mark" button via the touchpad when they identified a speaker transition. Last, we asked for participants' preferences, challenges encountered, suggestions for the different contexts of use, overall experience using visual and Haptic Captioning during a post-study interview. The experiment took approximately 40-minutes per participant.

Results and Discussion of Study 1

Perceived Speaker Transition

We computed the overall average accuracy for speaker indication per captioning modalities in Figure 5. Haptic Captioning: (Mean=93.75, SD =25), Position Caption: (Mean=83.75, SD =20.94), Avatar Caption: (Mean=83.06, SD =35.45), Color Caption: (Mean=81.44, SD =26.24), Traditional Caption: (Mean=80.19, SD =20.36), Pointer Caption: (Mean=73.75, SD =17.46), Speaker-name Caption: (Mean=76.94, SD =30.77). We performed one-way repeated measures ANOVA on the accuracy of perceived speaker transition and found no significant differences ($F=1.208$, $p=0.309$).

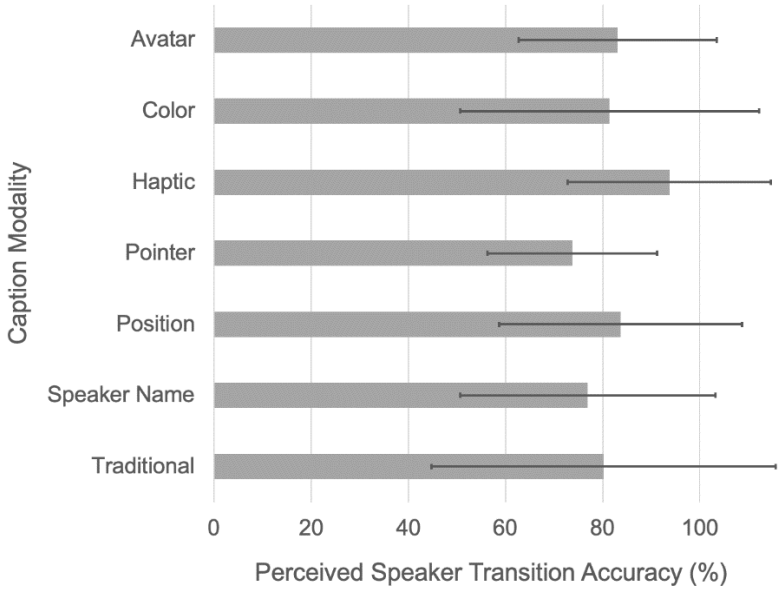

Figure 5: The average accuracy (percentage) of perceived speaker transition. From the highest to the lowest: Haptic Caption, Position Caption, Avatar Caption, Color Caption, Traditional Caption, Speaker-name Caption, Pointer Caption

Overall, the Haptic Captioning condition achieved the highest overall accuracy for detecting speaker transitions. Although this is surprising, some participants indicated that it may have been 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 brings focus to the usability of the Haptic Captioning system if there are frequent overlapping conversations. Besides the overlapping of voices, Haptic Captioning may also be affected by other audio content such as background music, audience clapping, etc. as all such audio signals would be provided as haptic feedback. P4 commented below.

---

*"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. In comparison, visual captioning such as pointer captioning, the switch between speakers is easier to identify even if the number of speakers is more."- **P4***

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 [32]. 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 are 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 have also been noted in previous works that investigated visual captioning styles [32]. Furthermore, P9 indicated *"not all of the colors are friendly to use for those who are color blind".*

Participant Preferences

Based on the result of user preference, the majority of the participants preferred visual-based captioning systems with the highest preference being for the Avatar Caption (8 out of 16) and followed by Position Caption (5 out of 16) method. Here, it should be noted that the visual captions that we presented were pre-processed videos with the caption styles added prior to the study design. However, 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 does not rely on any automated or manual identification of the audio content and is capable of providing a real-time translation of the presented audio as haptic feedback. As such, the Haptic Captioning condition was preferred more over the traditional captions methods that are typically associated with real-time captions.

Although only one participant selected the Haptic Captioning system as a favorite when asked to compare Haptic Captioning system with visual captioning methods, several participants expressed their positive feedback. For example, P7 commented that haptic feedback is better than other modalities since the vibration added a pause between speakers as *"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) mentioned *"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."*

Many participants mentioned that the Haptic Captioning system would be more beneficial by making it compatible with the 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 focus and enjoy more on the visual aspect of the media rather than paying attention to reading the text.

> *"Haptic Captioning methods experience a uniqueness in comparison to watching videos through visual captioning methods. In other words, Haptic Captioning allows me to not focus on text so much in video, and appreciate the visual aspect more." –* ***P12***

Some participants (P2\&P7) also expressed their desire to experience a longer time using the Haptic Captioning system with movies, music, and other genres.

"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." – ***P2***

"…I wanted to wear the haptic devices what I have a plan to watch any movie and see an ambiguous caption with the help of the haptic devices." – ***P7***

Summary of Study 1

In this study, we compared the Haptic Captioning system with several other captioning methods. Overall, while the Haptic Captioning system was rated lower in the participant preferences, participants provided many positive comments regarding our method. Here, one main suggestion was to explore the system with different types of content and applications. Therefore, next, we further investigate the user experience of applying the Haptic Captioning system in several scenarios with various media sources.

7 Study 2: Contextual Interview

Inspired by the feedback from the previous studies, we aim to explore DHH users' feedback on using the Haptic Captioning system different application settings. Therefore, we conduct a semi-structured contextual interview to answer RQ3.

Study Design

This study aims to understand DHH participants' experience with using the haptic modality in different contexts of using captions (i.e., TV, mobile, laptop). We designed a contextual interview to understand the user preference and then explore the possibility of the future Haptic Captioning system designs in terms of the context of use. Informed by Study 1, we made three video clips where each clip consisted of videos of four genres of videos (i.e., podcast, sports, live stream, movie). Each video is approximately 4 minutes (1 min * 4 genres). Although participants in study 2 suggested testing the haptic device for music, we did not include music in this study, considering the goal of speaker identification. Based on suggestions from the previous studies, we also asked the participants to try out the Haptic Captioning on three locations (i.e., wearing on the wrist, holding/attaching against the phone, attaching on the chair) and select the preferred location. In order to discover more comfortable positions, participants were allowed to change the vibrating position during the study. The order of the three settings was counterbalanced.

Participants

We recruited six participants (three male, two female, one non-binary) aged 18-26 (M = 22, SD = 3.4) from social media and the institute's mailing lists. 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 none. Two participants preferred to communicate through oral communication, and four participants chose to type in their feedback. According to the demographic questionnaire, all participants reported they had used captions on TV, mobile phone, and laptop. For the familiarity of three settings, TV: three participants were extremely familiar (R1, R4, R5, R6), one moderately familiar (R3), and one slightly familiar (R2). All participants reported being extremely familiar with mobile phone and laptop settings.

Study Procedure

In study 3, each participant experienced using the Haptic Captioning system on three devices with a four-minute video containing four media types genres. Before it started, participants were asked to select the vibrating position they felt comfortable using. We provided three positions as suggestions, but participants are free to change as they see fit: wearing on the wrist, holding against the phone, and putting on the chair. Last, we conducted a semi-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 preference and how the environment affects their experience. In the end, participants provided their suggestions to improve the system device, if any. In total, the contextual interview took approximately 50-minute per participant.

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 for searching emerged themes. We identified three themes which are **overall experience, vibrating position,** and **usage scenario**. We will present our major findings using the inductive themes and representative quotes below.

Overall experience

Overall, participants reflected their experience of using the Haptic Captioning system remains positive as it enhances their experience of watching videos by bringing more sense, especially to watching movies. Participants mentioned they enjoyed using the system and were more engaged without any distraction. Participants mentioned, *"I found it to be a nice additional dimension to the media"* (R1), and *"I think the movie will be a good experience using the Haptic Captioning system because it provides better senses."* (R6)

Enhance understanding emotions. Beyond the positive experience, 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 is helpful for matching the actions that happened to the sound effects, which are hard for them to access. This observation is an indication that the Haptic Captioning system could provide feedback on non-speech information.

---

*"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." – **R3***

Improve caption readability. Participants' responses also indicated that the Haptic Captioning system is beneficial for assisting the caption readability. For example, R3 mentioned that haptic feedback helps the captions as a supplement by matching the textual input of words with the movement from lip reading. Thus, especially when people can't 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." – **R3***

Assist speaker indication. While watching the video without seeing the picture of speakers, such as in a podcast, participants mentioned that haptic feedback helps identify the active speaker depending on how the voice vibrates. As participants perceived a female voice felt softer while the male voice seemed deeper in general, which aligned with the feedback of the Preliminary Study. However, participants also mentioned it's challenging to pick up the speaker's voice when sound quality is low.

*"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."– **R3***

Overall, participants found the Haptic Captioning system helpful for speaker indication, but they commented on a few difficulties recognizing who the speaker switched to, such as due to bad quality of the audio. Low quality can also affect poor captions, which can also contribute to a negative experience. However, participants indicated that with more training, they would be more confident in understanding different vibrations; as R2 commented *"Once I get more used to it, I don't feel it was challenging. I think it's more beneficial. Like, the more I use it, the more I understand like different vibrations. "*

Vibrating position

From the observations, we identified four vibrating positions that the participants preferred (Figure 6). With the Phone: three participants chose to wear it on the wrist (R1, R2, R6), and three held it against/attached to the phone (R3, R4, R5). With the TV: three participants preferred wearing it on the wrist (R1, R2, R4), one participant put it on her leg (R5), one participant attached the device to the phone (R6), and one participant first held it on hand (R3). With the Laptop: 5 participants preferred to wear it on the wrist (R1, R3, R4, R5, R6), and one participant (R2) first wore it on the wrist and switched to holding it on the hand to the lower vibration intensity in movie and sports.

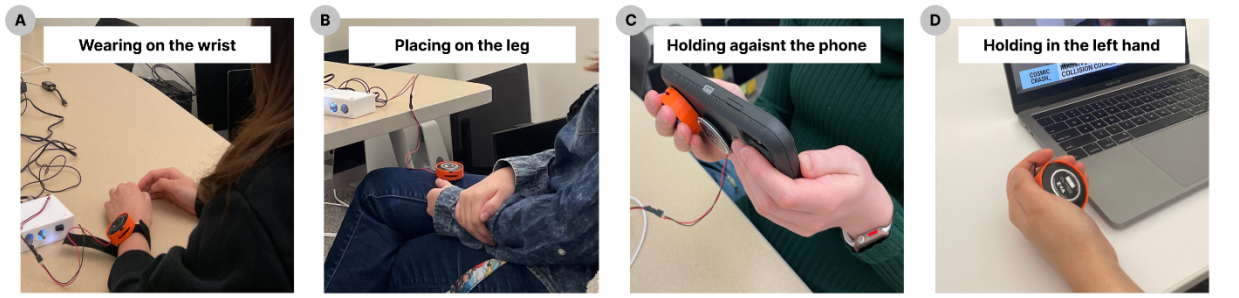

Figure6: From the left to the right: four vibration positions that participants prefer to use: wearing on the wrist, placing on the leg, holding against the phone, holding in the right hand.

Attaching to the phone. Participants generally prefer using the Haptic Captioning system on the phone rather than on TV and laptop since the vibration amplifies the most on smaller screens with lower volumes (see in Figure 6 C). Besides, attaching the device or holding the device against the back of the phone allows users to feel the vibration through their fingers. Moreover, one participant elaborated on the experience of attaching the device to the phone and wearing it on the wrist. Participants could best feel the vibrations when holding the phone with the device in both hands.

> *"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. Like I couldn't really focus on what they are saying. It's like if only one haptic on the right hand, I feel not completely connect with the left hand too. It got me somehow to miss some parts. However, on the phone with my both hands, I can understand what they are saying." – **R4***

Holding might not always be useful if participants want to multitask like having snacks while enjoying a video; as R1 mentioned *"Holding it was fine for the phone but would be annoying for laptop or TV."*

Wearing on the wrist. The majority of participants think wearing the Haptic Captioning device on the wrist is preferable (Figure 6a). This observation also aligned with a previous survey on DHH user preference on wearable devices, as 75% participants would prefer to receive haptic feedback from a separate wearable device [12]. Several participants mentioned their hands are more relaxed by wearing on the wrist, particularly after holding it for a long time. For example, R2 highlighted that *"I get tired from holding that so. That's why I just like kind of relaxed my hands, so I think it's like easier to the position like this [wear on wrist]"*

However, participants mentioned they have less feeling through the bone. When asked about the future design, participants suggested making it more transportable, such as making it wireless, more comfortable to wear for a longer period, and providing the ability to adjust intensity.

Placing on leg/chair. We noticed that in the TV set where we suggested participants place the vibrator on the chair first, they would prefer to change to other positions like placing it on the leg shown in Figure 6 B. Later, this participant (R1) explained that the vibration of the device was weaker on the leg than on the wrist but stronger than on placing on the chair. Participants further suggested testing the Haptic Captioning system built-in to a chair, which would specifically benefit the TV settings, as R1 commented below.

"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 wouldn't have to hold anything (in case you are signing with friends or something)." – R1

Usage Scenario

When asked about how the environment affects your experience, we identified several factors depending on different contexts, which will be explained below.

Environmental sound. As this experiment was carried out in a lab setting, few participants mentioned their sensations would be different than watching TV with a group of people. For example, haptic feedback might cause distraction in their conversation with others but 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" – R1

The "social acceptability" of Haptic Captioning indicated that 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 user preference on the wearing position. 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'd be more willing to kind of hold it against my phone. So it's less obvious. If I were in a private space, I'd 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." – R3

R2 further elaborated more on the difference between hearing in quiet and loud environments. For quiet settings, DHH users might rely on their hearing aids more than on reading the caption. However, he 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, now on 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. So like, why not? I thought it's I just keep hearing this. But in a loud settings, probably, probably better." – R2

Extra vibration from the environment might cause confusion on the understanding of the haptic information, such as multiple factors in the 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. – R4

Attention requirement. Most participants mentioned some context of using Haptic Captioning helps them to pay more attention, which could depend on the genres of the video. For example, participants mentioned for the podcast that the Haptic Captioning system helps enhance captions due to the lack of attention to visual details. Similarly, for the movie, where people sometimes do not focus on speaking much, participants feel more positive about the usage of the system.

" 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." – R2

Participants do not always want to focus on the content of media. Rather, they would like to watch the TV show without hearing aids. In this situation, the Haptic Captioning system helps them 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. – R3

8 Discussion

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

Takeaways from Three User Studies

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

Similarity between speakers' voice patterns. Our findings from three studies suggest that DHH participants found the difficulties of identifying speakers varied from the number of speakers and their background, which extends the challenge found in previous work [32]. In the Preliminary Study in 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. The quantitative results were also supported by participants' comments in the post-study session of the Preliminary Study and the studies 1 and 2. In addition, through the accuracy of identifying the demographic information (perceived gender, age), we noticed that the accuracy depends on the variety of the speaker demographic, especially their voice patterns. The accuracy is extremely low (less than 30%) when the speakers have a high level of similarity. In contrast, participants found it's easier to identify two speakers with distinguishable sound patterns (e.g., one male & one female or one elder & one child). Although we acknowledge this factor could relate to the recording quality, we suggest future design should amplify the difference between speakers' voices.

Familiarity towards devices and media. Participants' feedback in Study 1 revealed that the level of experience affects their preference in general, as P09 commented "*Well, all captions are hard to focus. It takes training (grew up with it) to become comfortable on where to focus.*" In study 2, one participant also explained that the level of familiarity with the haptic feedback affects his understanding. R4 commented that his hearing aids allowed him to be acquainted with the sound pattern of the environment. *"I assumed I'm used to the sound environment where I come from because I always wear hearing aids all the time, so the haptic feedback already affects my mind that I know what the sounds are like."* Similar factors related to the familiarity were also found in study 2. For example, few participants mentioned the relatively low frequency of watching TV compared to using a phone and laptop. Therefore, our next steps in this direction are to provide haptic training or design a longitude study to explore the Haptic Captioning in depth.

Attention required on the visual information. Our study extends on existing literature on 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 visual and haptic modalities. The comparison of haptic and visual captioning modalities revealed that while DHH people generally prefer visual caption modality, extra visual adds-on would raise new challenges on the increased eye fatigue and distraction, which impact the readability of

caption [20]. 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 text. 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.

Design Implication of Haptic Captioning System

Our findings indicate the future design of 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, strong vibration would also cause sound leakage, which some participants were worried about using the Haptic Captioning system in the public scenario. Therefore, participants should be allowed to customize the volume level as they see fit in the environment. While many works have attempted integrating haptic devices in such contexts and attached to mobile [23], we are motivated to explore this in a captioning context.

Second, 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 [29]. We aim to explore a wristband prototype that provides spatial haptic feedback with multiple actuators where the haptic feedback associates with the position of the speakers. It's 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 the build-in systems to provide a more immersive experience without occupying the 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.

Third, inspired by participant feedback, we aim 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 the speakers are speaking in an angry tone or sad tone, etc. While presenting non-speech information has been explored in the past with haptic feedback [22], we posit 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.

Suggestions for future study design

To measure the efficacy of the speaker indication, we designed an experiment interface that is able to collect the perceived speaker transition. This interface mainly contains a clickable “Mark” button which can be used when the participants feel the transitions between speakers. However, we noticed that in

some cases, participants marked the overlapping between speakers as the transition. To improve the interface design, future work should consider the difference between the perceived overlapping and perceived transition. In addition, participants mentioned by focusing on the experiment interface, it's hard for them to concentrate on the video content. Considering the cognitive load from the task itself, future studies could consider interface training at the initial part of the study or include the cognitive load examination as one index of the measurement. Another strategy for examining the factors of the task is to walk through the missed transition point with participants and debrief the reasons with them. The potential factors could include the experiment interface, the distraction from the caption, the distraction from the visual information, misunderstanding of the caption method, etc. Last, the speaker transition should not be the only method to examine the efficacy of speaker indication. Other dimensions such as engagement, ease to follow, distraction, and understanding should also be taken into consideration, which could be examined through eye-tracking or self-report questionnaires.

9 Limitations and Future Work

As a starting point, in the Preliminary Study, we only selected the clips with two or three speakers varied by the perceived gender and age. Therefore, we suggest future work examine the haptic perception with more than three speakers and address the challenge of identifying multiple speakers with similar demographic and sound patterns. We also acknowledge other factors of video clips might exist in Study 1, such as the total number of speaker transitions and the times off-screen shown.

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 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 preference.

Study 1 first examined the efficacy of the captioning methods through perceived speaker transitions. The speaker transition task itself might affect participants' preference as they need to concentrate on identifying the change. Future work can explore the captioning evaluation methods combined with the speaker transition task and the subjective questionnaire in terms of engagement, comprehension, and distraction. As we described before, few participants mentioned the fatigue of moving eyes constantly and the extra cognitive load of associating the visual indicators with the speaker information. Thus, we encourage future studies to investigate the cognitive load between haptic and visual captioning using questionnaires like NASA-TLX and eye-tracking.

Although we put efforts into setting up the contextual interview with three usage video watching scenarios (TV, laptop, phone), few participants mentioned the isolated lab could be a constraint which is different from the in-wild study. This is another direction we wish to explore in the future with the Haptic Captioning using in the different real-world scenarios, such as the environment participants are familiar or not familiar with, the public or private space, and the noisy scenario with external sound or not.

10 Conclusion

This study has investigated Haptic Captioning system through three-phase experiments. First, we examined the user perception of the speaker demographic with only haptic feedback. By analyzing the quantitative data, we found haptic feedback can convey the speaker's information, but the complexity would increase with the number of speakers. Further, through a within-subjects study with 16 DHH participants, we compared the Haptic Captioning system with existing visual modalities. While participants prefer visual captioning methods, there is no significant difference in identifying speaker transition between haptic and visual captioning methods. In the end, we conducted a contextual interview to understand user experience using Haptic Captioning system in three semi-realistic settings (TV, mobile, laptop) and observed participant preferences on wearing positions. Our qualitative data analysis suggested the overall characteristics of Haptic Captioning system and informed the future direction of design and research.

11 Acknowledgements

First and the most, I would like to give my warmest thanks to my advisor Dr. Roshan Peiris for the invaluable guidance and continuous support which make this work possible. I want to thank my committee Dr. Tae Oh for his encouragement and feedback during my master study and research. I also want to express my sincere gratitude to Pratheep Kumar Chelladurai for his assistance and other members in CAIR who gave their help during the period of this project. Finally, I am extremely grateful to my family and friends for all the support they have shared. They are my most important inspiration to complete this work.

12 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 18

[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] 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

[21] 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).

[22] 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

[23] 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.

[24] 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.

[25] 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

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

[27] 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.

[28] 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.

[29] 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.

[30] Yi-Hao Peng, Ming-Wei Hsi, Paul Taele, Ting-Yu Lin, Po-En Lai, Leon Hsu, Tzu-chuan Chen, Te-Yen Wu, Yu-An Chen, Hsien-Hui Tang, and Mike Y. Chen. 2018. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3173867

[31] 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.

[32] Quoc V ty 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.

[33] 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

[34] 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