Beyond Subtitles: Captioning and Visualizing Non-speech Sounds to Improve Accessibility of User-Generated Videos

Oliver Alonzo[*]
oa7652@rit.edu
Rochester Institute of Technology
Rochester, NY, USA

Hijung Valentina Shin
vshin@adobe.com
Adobe Research
Cambridge, MA, USA

Dingzeyu Li
dinli@adobe.com
Adobe Research
Seattle, WA, USA

ABSTRACT

Captioning provides access to sounds in audio-visual content for people who are Deaf or Hard-of-hearing (DHH). As user-generated content in online videos grows in prevalence, researchers have explored using automatic speech recognition (ASR) to automate captioning. However, definitions of captions (as compared to subtitles) include non-speech sounds, which ASR typically does not capture as it focuses on speech. Thus, we explore DHH viewers' and hearing video creators' perspectives on captioning non-speech sounds in user-generated online videos using text or graphics. Formative interviews with 11 DHH participants informed the design and implementation of a prototype interface for authoring text-based and graphic captions using automatic sound event detection, which was then evaluated with 10 hearing video creators. Our findings include identifying DHH viewers' interests in having *important* non-speech sounds included in captions, as well as various criteria for sound selection and the appropriateness of text-based versus graphic captions of non-speech sounds. Our findings also include hearing creators' requirements for automatic tools to assist them in captioning non-speech sounds.

CCS CONCEPTS

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

KEYWORDS

accessibility, audio tagging, automatic captions, non-speech sounds

ACM Reference Format:
Oliver Alonzo, Hijung Valentina Shin, and Dingzeyu Li. 2022. Beyond Subtitles: Captioning and Visualizing Non-speech Sounds to Improve Accessibility of User-Generated Videos. In *The 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '22), October 23–26, 2022, Athens, Greece.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3517428.3544808

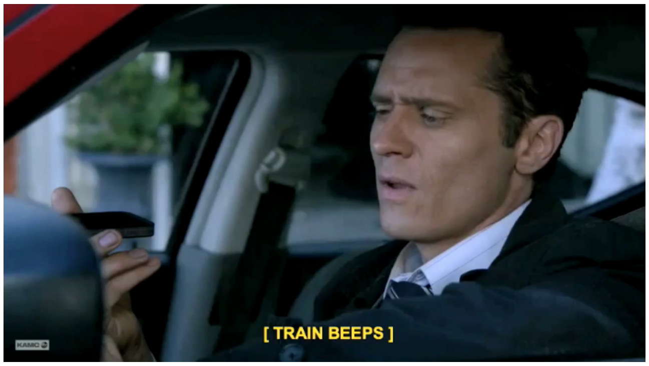


Figure 1: The non-speech sound [TRAIN BEEPS] is captioned to provide access to this sound for DHH viewers in a frame from the TV show *Castle*. (obtained from [35]).

1 INTRODUCTION

Closed captioning provides access to the audio of audio-visual content for Deaf or Hard-of-hearing (DHH) people . Using automatic speech recognition (ASR) to support automatic captioning of online videos has been increasingly explored, with several online platforms supporting its use (e.g. YouTube [34]). However, auditory content includes a richer array of sounds beyond speech such as music, background noises, or other non-speech sounds like laughter. Thus, automatic *speech* recognition alone is not enough to create complete closed captions that include such non-speech sounds [35].

To the best of our knowledge, the use of automatic sound event detection when captioning user-generated videos has not been explored (except for a blog post from YouTube [12]). Guidelines for manual or professional captioning include suggestions for including non-speech sounds (e.g. guidelines provided by 3Play Media[1], Web Accessibility Initiative[2], Described and Captioned Media Program[3] and BBC[4]). Thus, professionally-produced captions often include non-speech sounds. However, published research on automatic captioning for user-generated content mostly focuses on spoken content [35]. Thus, in this work, we explore the interests and perspectives of DHH adults on the inclusion of non-speech sounds in the context of user-generated videos. We also investigate how to support content creators in captioning non-speech information and sounds by using automatic sound event detection tools.

[*] This work was conducted during a summer internship at Adobe Research.

[1] https://www.3playmedia.com/blog/captioning-sound-effects-in-tv-and-movies/
[2] https://www.w3.org/WAI/media/av/captions/
[3] https://dcmp.org/learn/602-captioning-key---sound-effects-and-music
[4] https://bbc.github.io/subtitle-guidelines/#Sound-effects

ASSETS '22, October 23–26, 2022, Athens, Greece
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9258-7/22/10…$15.00
https://doi.org/10.1145/3517428.3544808

Guidelines for captioning non-speech sounds typically suggest the use of verbal descriptions enclosed in brackets (Figure 1). However, considering that non-speech sounds can also be visualized graphically, we also explore the use of graphic captions.

To this end, we conducted formative interviews with 11 DHH participants about their experiences with online videos, captions, and with non-speech sounds in online videos. We asked questions about their interests in having non-speech sounds captioned in online videos, including *which* sounds would be of interest, and *how* those sounds should be captioned (e.g. through text-based or graphic captions). Our results suggest interest in having *important* non-speech sounds captioned in online videos. How to caption those sounds may vary based on the type of video content, the sound type, and the intended audience. We also identified trade-offs between text-based and graphic captions.

Our formative study informed the design and implementation of a high-fidelity prototype to caption or visualize non-speech sounds using automatic sound event detection tools. We then conducted a study with 10 hearing video creators, asking them about their experiences creating online videos, adding or not adding captions, and about their thoughts on captioning or visualizing non-speech sounds, or how automatic tools can support them in this process. Then, they interacted with our prototype to caption and visualize non-speech sounds in three sample videos. Hearing video creators wanted automatic systems to be selective about the sounds identified and suggested that the appropriateness of graphic captions may vary based on the video content. Accurate time stamps and general descriptions were highlighted as important for automatic systems to provide.

The contributions of our work include:

1. Empirical evidence of DHH viewer's preferences for what non-speech sounds to caption and how to caption them, and hearing video creators' perspectives about what kind of support they want for captioning non-speech sounds.
2. Guidance for designers of captioning technologies and researchers in audio-visual analysis fields investigating technologies that may support the captioning of non-speech sounds in user-generated videos.
3. A high-fidelity prototype for captioning or visualizing non-speech sounds using automatic sound event detection tools.

2 BACKGROUND AND RELATED WORK

Captioning provides access to audio content as text and is often used to provide access to auditory content for DHH people [35]. While the terms *captions* and *subtitles* are often used interchangeably (e.g. [10]), their purposes may be different: subtitles display the *language* of the audio-visual content for people who do not know that language (e.g., a non-English speaker watching an English movie with subtitles in their language), while captions display the *audio* for people who do not have access to it (e.g., some people who are DHH). Thus, definitions of and guidelines for captioning include non-speech information and sounds such as speaker information, environmental noises and sounds, sound effects, music, etc. Captioning is legally required for content streamed on live TV in countries such as the U.S. [16], but no such requirements exist for online videos. Thus there is a vast difference in the availability and quality of captions across online platforms.

Research has explored various aspects of the user experience (UX) and personalization of *subtitles* (e.g. [10, 11, 14, 15, 19]), as well as the needs of diverse users (e.g. [1]. However, little work has focused specifically on the perspectives of DHH viewers and hearing creators on the inclusion of non-speech sounds using text-based or graphic captions.

2.1 ASR and Automatic Captions

Video authoring and communication tools, as well as online and social media platforms, are increasingly adopting ASR to support automatic captions (e.g. Premiere Pro[5], Zoom[6], YouTube[7], Instagram[8], TikTok[9]). However, research on the use of ASR for automatic captions is still on-going.

Prior work has examined the preferred *appearances* of captions among DHH adults, finding great diversity in preferences towards various visual characteristics (e.g. font, background color) [4], and preferences towards the use of punctuation in automatically generated captions [20]. Prior work has also examined how to evaluate ASR systems among DHH adults, finding that the literacy levels of participants (which are diverse among DHH adults [29, 30, 32]) affect the effectiveness of metrics typically employed for captioning evaluation [5]. More recent work explored which genres DHH adults prioritize for accurate captioning, finding that news and politics, education, tech and science, and film and animation were among the top priorities [6]. Finally, semi-automatic captioning approaches using crowd-sourcing have also been explored [31].

However, prior work on the use of ASR for automatic captions has mostly focused on *speech*. To the best of our knowledge, the only exception is a blog post from YouTube [12] which describes the inclusion of three non-speech sounds (music, applause, and laughter), but does not provide details about the user study supporting that decision. Thus, in this work, we explore the perspectives of DHH adults on the inclusion of non-speech sounds when using automatic tools for captioning user-generated videos.

2.2 Non-Speech Sounds in Real Life and VR

Research has explored using automatic sound event detection to create sound-awareness applications for DHH users in physical environments. Prior work includes investigations of which sounds are of importance, what aspects of those sounds are of importance, where sound awareness is more important, approaches to visualize sounds, and the appropriateness of various devices for visualizing sounds in physical environments (e.g. [7, 18, 22]). Findings from prior work reveal urgent and safety-related sounds as sounds of general interest for sound awareness [18]. However, participants' hearing ability may affect those interests [18]. Some characteristics of sounds, such as a sound's source, identity and location, may be more important than other characteristics such as its volume [18].

Recent work has also explored non-speech sounds in virtual reality (VR). Jain et al. developed a taxonomy of sounds as a starting

[5] https://helpx.adobe.com/premiere-pro/using/speech-to-text.html
[6] https://blog.zoom.us/update-on-live-transcription-for-free-accounts/
[7] https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html
[8] https://about.fb.com/news/2020/09/new-automated-captions-powered-by-ai/
[9] https://newsroom.tiktok.com/en-us/introducing-auto-captions

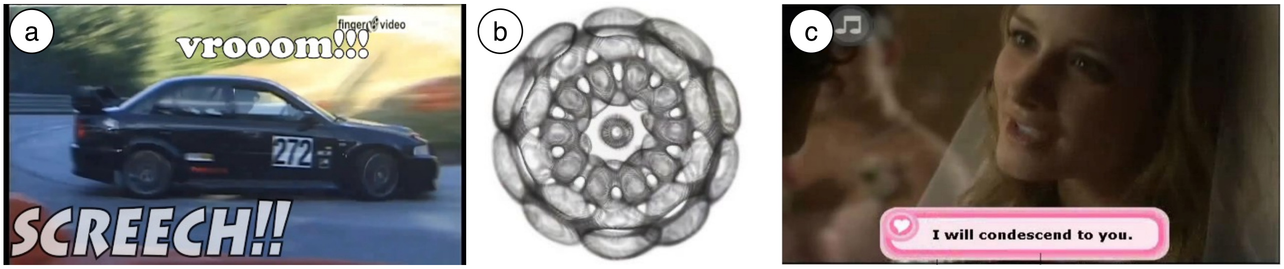


Figure 2: Examples from prior work on different visualization techniques, including: a) dynamic text that varies in size to indicate the volume of non-speech sounds [33], b) a physics-based approaches that indicate how sound would move through physical materials [27], and c) using colors and icons for visualizing emotion in spoken content [25].

point to explore sound awareness in VR based on two dimensions of sounds: their source and intent [21]. As the need for separate exploration of non-speech sound awareness in VR highlights, the findings from one domain (e.g., physical spaces) may not necessarily translate to others. However, to the best of our knowledge, the inclusion of non-speech sounds in the context of authoring captions for user-generated online videos has not been explored.

2.3 Visualizing Sounds

Sounds can be visualized in several ways that vary how they relate to properties of the sounds or their level of semantic meaning. As shown in Figure 2a, text with dynamic size has been explored to indicate volume when captioning non-speech sounds [33]. Prior work has also explored visualizing other non-speech information, such as emotion, using icons and colors (Figure 2c) [17, 25]. Researchers in [17] explored the use of icons and colors to augment existing speech-oriented text-based captions, which was described as potentially childish by participants. Furthermore, researchers have explored visualizations as a way to supplement tactile information when exploring the use of tactile feedback to provide non-speech information to DHH viewers, finding benefits when using both combined such as improved recall of non-speech information [24]. However, in that work, researchers did not explore participants' preferences for visualizations alone, as they were only explored alongisde tactile feedback.

Research has also explored physics-based approaches, such as cymatics (Figure 2b), which imitate the movement of physical matter among the waves produced by sounds [27]. These approaches can be useful in settings such as audio editing or visual explorations of characteristics of sounds. However, they are not semantically meaningful. Thus, prior research on applications of sound awareness [26] or visualization of non-speech sounds (e.g., in video games [13]) typically includes more semantically meaningful approaches such as icons to represent a sound's source.

GIFs and animated stickers, which can be overlaid on videos, have been growing in popularity on social media and may provide new ways for visualizing sounds in semantically meaningful ways that are separate from text-based captions (see Figure 3), but to the best of our knowledge, no work has explored their use as graphic captions for non-speech sounds. Thus, in this work, we also explore the use of such graphic captions as a potential alternative to text-based captions in the context of user-generated videos.

3 RESEARCH QUESTIONS

Based on the gaps identified above, in this, study we investigate the following research questions:

First, focusing on DHH viewers, we investigate:

- **RQ1**. What are the experiences of DHH viewers with online videos (e.g. what type of content do they watch, what do they like and dislike about it), and with closed captions?
- **RQ2**. What are DHH viewers' perspectives on the inclusion of text-based and graphic captions for non-speech sounds in online videos?

Then, focusing on hearing video creators, we investigate:

- **RQ3**. What are the current practices of hearing creators with captioning online videos?
- **RQ4**. How do hearing creators perceive the use of a prototype tool based on automatic sound event detection for captioning non-speech sounds using text-based and graphic captions in online videos?

4 INTERVIEWS WITH DHH PARTICIPANTS

To answer RQs 1 and 2, and inform the design of a prototype used for RQs 3 and 4 (described in section 6.1), we conducted a formative study with DHH participants. This section describes the study's method, participants, and results.

4.1 Method

We conducted semi-structured interviews with DHH participants, and our electronic appendix includes our full questionnaire. We began by asking about their experiences watching online videos, their experiences with captions in online videos, and their thoughts about having non-speech sounds captioned or visualized.

Then, we prompted participants with three videos illustrating text-based and graphic captions for non-speech sounds. We picked three videos that varied in: their format (i.e. landscape or portrait); their genre, which we selected from prior work identifying genre priority for accurate captioning in online videos (more specifically, we selected one genre from each priority level identified: sports, news and entertainment) [6]; and their source, which included YouTube (Figure 3a), BBC (Figure 3b), and TikTok (Figure 3c). Our video selection was not meant to illustrate all possible combinations of those three aspects. Instead, we wanted some diverse combinations to prompt participants to think about these different aspects.

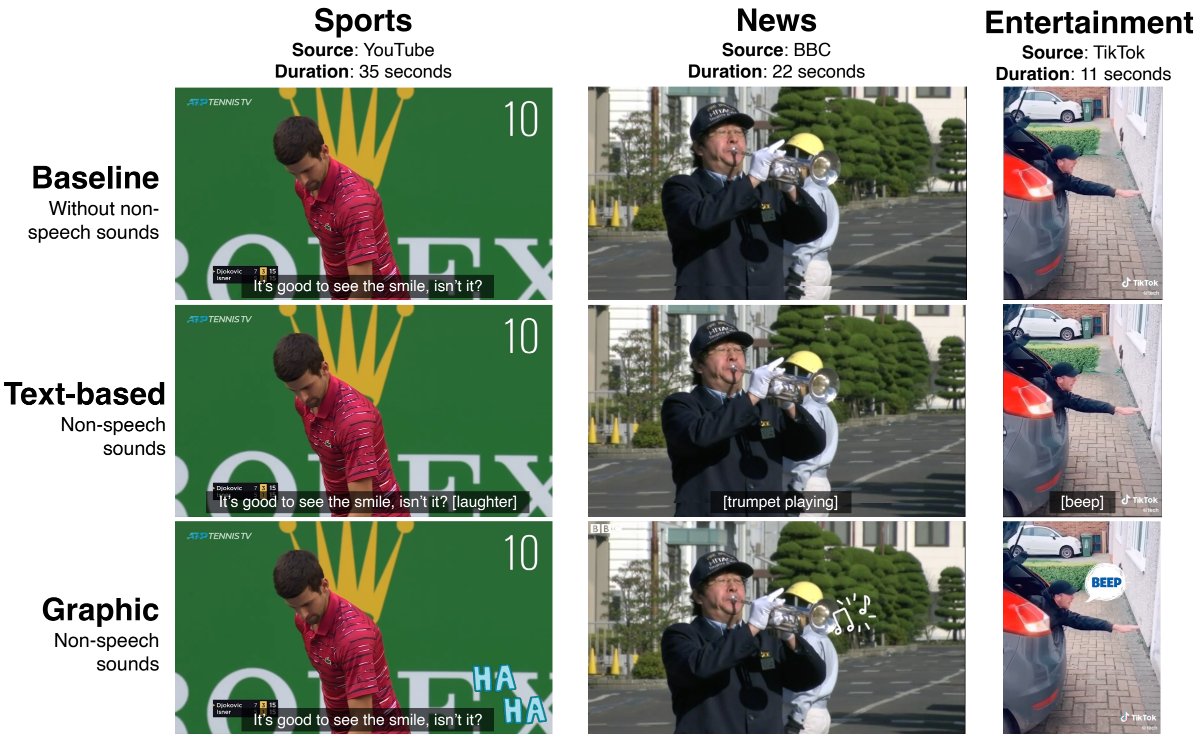


Figure 3: Frames from the video stimuli used in our studies. Each column illustrates the three videos, while each row illustrates the conditions the videos were shown in. The text-based captions have been enlarged for readability.

Thus, our three videos consisted of a YouTube video of a sports scene in landscape format, a BBC news video in landscape format, and a TikTok entertainment video in portrait format. More details about these videos are provided as part of our electronic appendix.

We showed each video to participants in three conditions (illustrated in Figure 3): a) without captioned non-speech sounds as a baseline, b) using text-based captions for non-speech sounds, and c) using graphic captions for non-speech sounds. The demonstrations for these conditions were created manually using Premiere Pro, using GIPHY[10] stickers for the graphic captions in condition c.

We always started with our baseline (condition a), and then we rotated the order of conditions b and c across participants. We also rotated the order of the videos using a Latin Square schedule. After watching all three conditions for each video, we asked participants what they liked or disliked about each version, about their perspectives on text-based and graphic captions for non-speech sounds, and whom they think these technologies may benefit most.

4.2 Participants

We recruited 11 participants through online advertisements posted on social media groups on Reddit and Facebook, including groups targeted for DHH people in the general population, and one group from a large university for DHH people in the U.S. Participants' mean age was 30 (range = 18 to 47, SD = 9.68), self-identifying as male (N = 4), female (N = 6) and non-binary (N = 1). Five self-identified as Deaf (Deaf, with a capital D, is usually employed to refer to members of Deaf culture [28]) and six as Hard-of-hearing.

4.3 Procedure and Data Analysis

Participants received a consent form via e-mail before the study and met with a researcher via Zoom. The appointments lasted 55 minutes on average. In the end, participants filled out a demographics form and were compensated with a USD$30 Amazon gift card.

We conducted the interviews in English, accommodating participants' self-selected communication preferences, which included using American Sign Language (ASL) interpreters (N = 3), professional captioners (N = 1), automatic captions (N = 3), text-based chat (N = 3) or spoken English alone (N = 1).

The transcripts, obtained using Zoom's automatic transcription, contained 4770 words on average. In the interviews supported by ASL interpreters, their voicing of participants' signing was transcribed. The first author then conducted a thematic analysis using an inductive approach to identify codes which were then grouped into themes, as described by Braun and Clarke in [8]. We also follow the best practices for reporting our findings as discussed in [9].

5 FORMATIVE INTERVIEW RESULTS

5.1 Online Videos and Captions

The most commonly mentioned video platform was YouTube. Others also mentioned watching videos on social media platforms,

[10] https://giphy.com

including Instagram, Facebook, and TikTok, as well as streaming platforms, such as Netflix, Amazon Prime, Hulu, HBO Max, and Disney+. Participants cited three main purposes for watching videos online: entertainment, educational, and informational purposes.

5.1.1 Participants like the control they have in online videos, but there is a lack of high-quality captions. Control was an aspect participants liked about online videos, including control over what to watch, the playback, volume, and the ability to turn captions on or off. For instance, P11 said "*When I'm watching videos online, I can pause it and if I miss something, I can rewind it. I also have control of the volume.*" Some also highlighted having visual information as a benefit, and how captioning may be beneficial in learning vocabulary. Participants also highlighted how captioning technologies for online videos, including automatic captioning, have improved overall.

However, captioning issues were prominent among what participants did not like about online videos, including inaccuracies in automatic captions and the difficulty of finding well-captioned content. P3 mentioned that well-captioned content may be expensive, citing subscription-based streaming platforms (which often caption non-speech sounds, according to P3) as examples. Finally, participants also mentioned the loss of captions when copying videos (e.g., videos reposted to social media), and the lack of support for captions in live videos (e.g., Twitch) or social media platforms.

5.1.2 Participants sought workarounds to understand uncaptioned content, but were mindful of others' experiences. Participants reported several workarounds for uncaptioned content such as asking someone, trying to understand by re-watching the videos, or finding the same content in writing. Some hard-of-hearing participants also indicated buying headphones or speakers to play videos louder, including P11 who said "*I put up a lot of volume in the sound bar, like I spent extra money to get extra sound.*". Some Deaf participants also mentioned reading people's lips. However, many participants indicated often leaving videos if one of the methods above fails, especially if a person is clearly talking in the video. For instance, P6 said "*If I encounter a video that I need to rewatch or go back several times, I'll probably just move on to another thing.*" Notably, P4 and P11 mentioned not wanting to affect others' experiences by asking them to interpret the content or by making sounds louder. For uncaptioned non-speech sounds, in addition to the workarounds mentioned above, participants indicated recognizing missing sounds through the behavior of people in the videos, or by noticing a gap in the story. While some mentioned instinctively recognizing that something is missing, others also acknowledged that they may still inadvertently miss non-speech sounds, including P1 who said "*I may miss out what happens beyond what people say without my knowledge.*".

5.2 Text-based vs. Graphic Captions.

5.2.1 Advantages and disadvantages of graphic captions. Many participants liked the ability of graphic captions to provide more details about a sound such as its source, location, tone of voice, volume or emotion. Participants found graphic captions easier to understand and see, which may make them more universal and benefit viewers who cannot read text-based captions (e.g., children), with some finding the visual nature of graphic captions more "ASL friendly." However, participants worried about their potential for distracting, blocking the content of the video, or taking away from the experience of watching online videos by impacting their emotional or visual feel. Some also worried that they may feel childish and annoy adults as it may feel like "talking down" to them. Participants sometimes found the graphic captions hard to distinguish from other graphics or visual effects in videos. Finally, most participants suggested that graphic captions should be provided on demand, giving viewers control over whether they see them.

5.2.2 Advantages and disadvantages of text-based captions. Participants liked the *familiarity* of text-based captions, including their location in standard places in the video, and the symbols used to indicate non-speech sounds (i.e., brackets). Participants liked that they "do the job" without disrupting the video. Participants saw the potential for text-based captions to include aspects of sounds that they liked that graphic captions included (e.g., changes, source, source location, and timing). For instance, P5 mentioned that text-based captions could use verbs to indicate changes (e.g. [applause fading]), P10 suggested that arrows could show where the sounds come from, and P3 suggested using additional symbols to layer these details.

However, participants mentioned that while text-based captions make videos accessible, they are often not interesting and lack context. Furthermore, an emergent theme was that it may be hard to verbally describe sounds for people who do not know those sounds. For example, P11, who identifies as hard-of-hearing, shared an anecdote of having to explain what "[roaring]" meant while watching a movie with Deaf friends. While P11 indicated people are now more familiar with brackets signifying non-speech sounds, they still worried about the difficulties of describing sounds with words, and suggested the creation of a glossary to support this.

5.2.3 How to choose between text-based vs. graphic captions. Various factors related to the videos, the sounds and the viewers were discussed as criteria to select between text-based and graphic captions. First, three aspects of a video were discussed, including its type, length, and how visually busy it is. Specifically, participants found graphic captions more appropriate for entertainment videos (e.g., in social media), whereas text-based captions appeared more fitting for more serious content. For instance, P6 said "*If I'm on social media [a] balloon is totally appropriate, right? But if I saw that same balloon representation in like BBC, not so much.*" Graphic captions also seemed more appropriate for shorter videos, whereas text-based captions may be better for visually busy scenes as well as videos that have spoken content. Second, participants discussed specific aspects of the sounds, including the visibility of the sound's source (i.e,. whether it is on screen) and dynamicity (i.e., whether a sound changes). Participants had mixed opinions about which was more appropriate for visible sounds, but graphic captions seemed better for dynamic sounds (e.g., an applause visualization decreasing in size as its volume decreases). Finally, participants brought up two aspects of viewers as criteria: their age and hearing abilities. P11 suggested that graphic captions seemed more appropriate for younger audiences, while P9, for instance, suggested that text-based captions may be better for people with slight hearing loss who, as P11 put it, may already have an understanding of the sounds.

A few participants also suggested that graphic and text-based captions could be used simultaneously, with one complementing the other, which could have the benefit of standardizing the location of the graphic captions if they were placed around the text-based captions to provide complementary information, or as P6 said, it would all be "*located in the area that I'm expecting to see that information*". P5 also suggested using emojis for "textifying"the captions (i.e., making them closer to the informal language commonly used in text messaging aided by emojis to convey emotion), which could emphasize emotion. Other participants, in turn, expressed indifference for the format, as long as the information is provided, such as P4 who believed that "*access is the key.*".

5.3 Sounds and Information to Include

5.3.1 Caption sounds that are important to the storyline of the video, and consider viewers' preferences. Participants expressed strong interest in having non-speech sounds included when captioning online videos. Most participants conditioned their preferences for what non-speech sounds to include on whether the sound is *important* to the video, as opposed to specific types of sounds. As P4 put it: "*specifically if it's relevant to the story. So for example, if a guy is going into a house and he's not talking but he hears something in the house, then that's important. But if for example, a woman is drinking water and you can hear her swallow, that's probably not relevant.*" This selectivity was in part motivated by the fact that there can be too much going on, especially when a video also contains spoken content. Other criteria related to the content of the video such as the number of people talking and what they're doing, and the length of the video, are also important to consider when selecting which sounds to visualize. Specifically, participants mentioned that as the number of people talking in a video increases, or if the level of activity of activity of people in videos increases, the need for being selective becomes more important. P9 also suggested being able to select only the sounds that are outside of a viewer's hearing range such as low frequencies (e.g., bass sounds) in the case of P9. Some participants also mentioned how some frequencies may interfere with their hearing aids and thus it may be useful to have them captioned (and perhaps silenced). Notably, many participants talked about the diversity within the DHH community, often referring to it as a "spectrum," and how this diversity may be reflected in their preferences for what sounds should be included. One participant, for instance, mentioned not liking music because they felt music is not compatible with Deaf culture. Thus, what counts as "important" sounds may have a high level of user-dependency.

5.3.2 Including details about the sounds is important. When talking about what aspects of sounds to include, participants highlighted details such as the source, its location, and temporal changes within the sounds. P4 also mentioned it is important to know how sounds interact with people in videos by detailing who can hear the sounds, for example. However, P11 also mentioned it is important to balance the level of detail and the length of the descriptions, especially for text-based captions, to consider slow readers and the fast-moving nature of captions. Finally, P2 highlighted the importance of details especially for members of Deaf culture, for whom details and descriptors are not only important, but can also be useful for learning vocabulary and learning about sounds:

> "*Because in Deaf culture, we usually like descriptors. We like to understand what vocabulary is applied to certain sounds so that if it's explained we can understand what's happening in English. So, that helps us as individuals to understand what hearing people are hearing. The same can be said, same could be true about Deaf people explaining our experience to hearing people, they may not understand so we have to be more detailed.*"

5.4 Benefits of Captioning Non-Speech Sounds.

Participants commented that the inclusion of non-speech sounds would benefit DHH people, as well as people with auditory processing disorders. Some also highlighted the importance of considering Deaf-blind people when incorporating these into videos, especially with graphic captions. Participants also reflected on how captioning non-speech sounds can also improve the co-viewing experience of DHH and hearing people, since DHH people would not need to ask hearing people for clarification to understand content.

Ultimately, the benefits of captioning non-speech sounds included understanding or knowing "what's going on" in videos, knowing the whole story, or "being at the same level" as everyone else. Participants indicated how sounds can convey information, set up scenes or the tone for a scene, give more depth, and provide spatial awareness, with P6 describing sounds as "condiment" for videos. Some participants commented how these may be realized subconsciously and thus seem unimportant to hearing people, but these functions are important for fully understanding "what's going on." Finally, P2 highlighted that text-based captions of non-speech sounds may help distinguish captioned videos with no dialogue from uncaptioned videos. In other words, text-based captions of non-speech sounds in a video without spoken content can indicate that the video is captioned (whereas an uncaptioned video may or may not contain spoken content). However, graphic captions alone may not have this same function as it may be hard to distinguish graphic captions for sounds from general visual effects in videos.

5.4.1 Tools that support captioning non-speech sounds may benefit hearing and DHH creators, as well as DHH viewers. While many participants indicated that tools to support including non-speech sounds as text-based or graphic captions would support hearing creators in making their videos more accessible, some also saw potential utility for DHH creators who would like to accessibly include non-speech sounds in their videos, but cannot hear those sounds themselves. Some participants also envisioned how the inclusion of non-speech sounds in fully automatic captioning would be useful for viewers, as viewers often do not have the choice to make videos accessible themselves when creators choose not do so.

6 PROTOTYPE STUDY

Based on the results from our formative interview study, we designed a second study to investigate how creators can be supported by automatic tools to caption or visualize non-speech sounds. To this end, we developed a prototype for authoring text-based and graphic captions of non-speech sounds for videos, and a video demonstration of this prototype is included as part of our electronic appendix. We conducted semi-structured interviews where participants interacted with our prototype and answered questions about their experiences before and after their interactions. This section describes our prototype, followed by our interview method, participants and data analysis.

6.1 Prototype

While we drew inspiration from existing captioning interfaces for our general interface (e.g. YouTube and Adobe Premiere Pro), our design was mostly function-driven, guided by the need to author (section 6.1.1) and preview (section 6.1.2) the captions. Moreover, our formative interview results also informed specific design decisions as highlighted below.

To use the prototype, users start by loading a video into the authoring interface. Our system displays a list of non-speech sounds in the video as would be detected by an automatic algorithm (Figure 4a), as well as a preview of the video (Figure 4g). Each occurrence of a non-speech sound is represented by a card user-interface element, which includes a text description of the sound, a time-stamp of when the sound occurs, and optionally a graphic that can be used as a graphic caption for the sound (Figure 4a). Motivated by the findings in our formative study, if the sound event card is generated by an automatic algorithm, it also includes a label suggesting the user to consider adding more details to the suggested description.

6.1.1 Authoring non-speech sound captions. Each sound event card is added as a caption to the video. The text description is used for the text-based caption, the graphic for the graphic caption, and the time stamp indicates when the caption will appear in the video. Users can add, delete or edit the sound card to customize the captions. They can change the text description or the timestamp by simply typing over the corresponding fields. To change the visuals, a user can click on "Add visuals," which opens a modal search box. The search box, powered by the GIPHY API for developers[11], supports searching for three types of visualizations: GIFs (animated images), Stickers (animated vector drawings or image cutouts) and Emojis (more specifically, stickers containing emojis). Users can specify the location, size and rotation of the visuals through direct manipulation in the video player while previewing the results. Finally, as participants in our formative study indicated that they were interested in *important* sounds, our prototype supports marking a subset of sounds as important by toggling the star icon on the top right corner of a sound event card.

6.1.2 Previewing the captions. Users can preview the captions in the video player by selecting the "Descriptions" option under the video player. Similarly, to preview graphic captions, the user selects the "Visuals" option (Figure 4b). By default, all sound event cards are included as captions in the preview. Users can select *Starred*, to only preview captions for the starred sound events.

6.1.3 Implementation. We built the front-end of the prototype using the React JavaScript library for creating the HTML components and functionality, and the Bootstrap library for the styling of the components. As noted above, the visualizations provided were obtained using the GIPHY API for developers, which provides options for requesting GIFs and Stickers. The Emoji search in our prototype was thus a search for stickers with the word “emoji” appended.

6.2 Method

The study began by asking participants about their experiences creating online videos, including how often they create videos for posting online and what types of videos they create. We then asked participants about their experiences captioning videos, which included questions about how often they caption their videos, how they make the decision to caption or not caption their videos, as well as what tools they use and what content they consider when they do caption their videos. We then asked them questions related to captioning non-speech sounds, including their current considerations of including non-speech sounds in captions.

We then introduced the prototype to participants through a 3-minute demo that introduced the video player, the sound events tab, and all the functions of the prototype. Then, participants interacted with the prototype to caption or visualize the three videos we had used in the formative interview study, as described in section 4.1 under three conditions: 1) only adding sound events manually (i.e. without using the Wizard-of-Oz automatic sound detection system); 2) a Wizard-of-Oz condition using results that were created by a member of our team to produce error-free output; and 3) using the direct output from an actual automatic sound event detection system [23], which contained errors. This structure allowed us to prompt participants for comparisons between manual additions and the Wizard-of-Oz automatic system, but also between human-quality output and automatic results containing errors. The automatic results included errors such as missing sounds, mislabeling sounds and labeling a sound multiple times. However, the time stamps for the sounds that were detected were often close to the actual timing of the sounds. All the labels included in these conditions and their time stamps are provided in our electronic appendix.

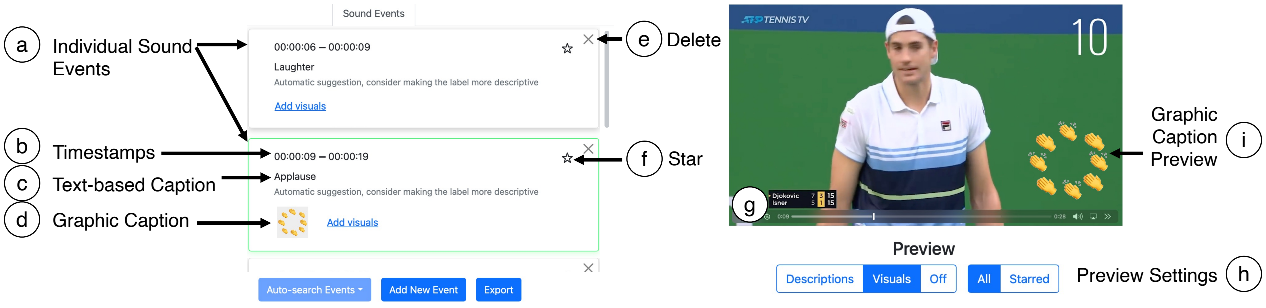
All participants interacted with the prototype in all three conditions. We asked participants to do the manual-only condition first to ask them to imagine what they would want from an automatic system to support them in this task. We rotated the other two conditions using the Wizard-of-Oz automatic systems across participants. We also rotated the videos using a Latin Square schedule.

We asked participants to think out loud as they captioned or visualized non-speech sounds in the videos. Once participants were satisfied with the results, they were exported as JSON files containing metadata for the text-based and graphic captions participants added so that we can replicate their work. After finishing each video, we asked about their experiences captioning or visualizing the non-speech sounds for each video. After finishing all three videos, we asked participants to reflect on their overall experiences, and to compare the use of text-based versus graphic captions as well as manual captioning versus automatic systems.

6.3 Participants

Participants were recruited from two sources: internal communication channels at an industry research lab, as well as special-interest social media groups, focused on video creation and university groups local to one of our co-authors. Our recruitment criteria included having experience creating videos for posting online, but we did not specify specific levels of experience so that we could get diverse levels and thus a diverse set of perspectives.

[11] https://developers.giphy.com



00:00:06 – 00:00:09
Laughter
Automatic suggestion, consider making the label more descriptive

00:00:09 – 00:00:19
Applause
Automatic suggestion, consider making the label more descriptive

Figure 4: Our prototype provides an interface to manage individual sound events (a) with their respective timestamps (b), enabling authoring text-based captions (c) and graphic captions (d). The user can highlight (f) or remove (e) these sound events. A a video player allows users (g) to preview the captions by selecting the appropriate setting (h) for text-based ("Descriptions") or graphic ("Visuals") captions. The figure shows "Visuals" selected and thus graphic captions are previewed (i).

We recruited ten participants, which included six who self-identified as male, three as female and one as non-binary. Participants' average age was 26 (range = 18 to 29, SD = 6.6). Participants self-reported levels of experience included developing (N = 3), competent (N = 5), advanced (N = 1) and expert (N = 1). Participants reported captioning the videos they create rarely (N = 3), occasionally (N = 2), often (N = 4) and always (N = 1).

6.4 Procedure

Participants were contacted via e-mail to receive a consent form ahead of the study. Participants then met with a researcher via Zoom for a 60-minute appointment. At the end of the appointment, participants filled out a demographics form and were compensated with a $20 Amazon gift card for their participation.

6.5 Data Analysis

The video recordings of the interviews were an average of 55 minutes long. Zoom's automatic transcription system was used throughout all the interviews to obtain transcripts of these videos, obtaining an average length of 7680 words per transcript. As we only collected open-ended comments from participants, our analysis was similar to our formative user study. Namely, the data from this study was analyzed by the first author using a thematic analysis approach, first using an inductive approach to identify codes in the data which were then grouped into themes following the approach described in [8, 9].

7 PROTOTYPE STUDY RESULTS

In this section, we summarize the results from our prototype study, which address RQs 3 and 4.

7.1 Previous Captioning Experiences

We had a mix of participants who either caption their video content regularly or rarely. Participants discussed motivations for captioning content, which included personal experiences benefiting from captions for comprehension or for understanding people with different accents. While some indicated only captioning videos when it is required for work or academic assignments, others were motivated to do so for ensuring their videos are accessible, especially when making videos for classes with classmates with cognitive disabilities or who are DHH. P5 also mentioned adding captions because of stylistic purposes to add personality (e.g. to emphasize something with specific fonts).

7.1.1 Participants tend to produce higher quality captions when the stakes are higher. Some participants indicated producing high-quality captions only for academic- or work-related videos. Participants indicated manually captioning videos, starting from an automatic system and editing the output, or paying an expert to do it. Most participants indicated relying on automatic systems to do some of the work, or an expert to do it for them, because of how tedious the manual captioning process is. For personal videos, however, many participants indicated using automatic captioning tools alone, even when they may be inaccurate. For instance, P1 said "*At work, the stakes are higher [...] whereas for personal videos it's a little more of a laid back approach [...] I'm doing things on YouTube and if I'm only going to get 10 to 100 views, then the auto captions are more than enough.*"

7.1.2 Most participants only captions videos with spoken content. Many participants indicated only captioning videos that have spoken content. For example, P4 said "*I'd say if there's spoken content, I'll do closed captions.*" Some participants mentioned *not* captioning videos when a video only contains music, text or pictures, or if there is no spoken content. Most participants indicated never having included non-speech sounds when captioning videos (which means they were creating, by definition, subtitles). However, 3 participants indicated having occasionally included them, including P5 who said: "*I wouldn't say consistently, every single time I encounter non-speech sounds, but I have captioned them before.*" P9 also mentioned adding a song's title if a video only contains music, or information about non-speech sounds if they are the main point of a video. P9 has also uploaded videos to social media platforms that do not support manually adding captions, but found workarounds to caption non-speech sounds such as commenting, replying or using a text overlay or a sticker.

7.2 Perspectives on Text-Based and Graphic Captions

The themes we identified related to hearing creators' perspectives on text-based and graphic captions included challenges for creating descriptions of non-speech sounds for text-based captions, as well as the appropriateness and benefits of graphic captions. The following subsections summarize these themes.

7.2.1 Challenges for creating text-based captions for non-speech sounds included trade-offs between completeness and concision. Many participants expressed doubts about the best way to describe sounds for text-based captions, the best wording, and what level of detail to include. Participants discussed trade-offs between length, complexity, accuracy, and how interesting they could make the descriptions. For example, P6 said "*I'm trying to think of how to put that across in a short, interesting way.*" Considering these trade-offs, participants commented that on-screen sounds may need fewer details as the sound source is already visible. For instance, P5 said: "*I think ['trumpet'] gets the point, especially if there's a guy playing the trumpet so I might leave that just for conciseness.*" Describing ambiguous sounds (i.e. sounds with unclear sources) was also challenging for some participants. For example, when describing a beeping noise from a human in one of the sample videos, P8 expressed confusion by saying: "*I didn't understand how to describe it. So, it was a bit confusing but I thought car beeping noise would be the most accurate description because I think everybody knows about a beeping noise.*"

Considering these challenges, some participants indicated that guidance for creating text-based captions would be useful. P4 indicated that for the text-based captions, guidance could be structural (e.g. using verbs vs. nouns, or what number of words to use). P1 and P5, in turn, suggested that because coming up with good descriptions may require a wider English vocabulary, having some support for description alternatives may also be useful.

7.2.2 The appropriateness of graphic captions varies with the type of video and certain characteristics of scenes. Most participants suggested that the appropriateness of graphic captions varies depending on the type of content. For example, most suggested that the visualizations available in our study were not appropriate for the BBC video because those tend to be more "serious." However, the visualizations seemed appropriate for the TikTok video. Both P2 and P3 specifically described the appropriateness depending on the "place and purpose" of the video. There was disagreement, however, about whether the formality of the visualizations themselves would affect their appropriateness. P2, for example, suggested that if the BBC created their own "more formal looking" set of visualizations, graphic captions could potentially be more appropriate. However, P5 indicated that "*even if we were to have more formal visuals for something like the BBC, I think it's just not in their guidelines to use [visual] effects.*" Finally, participants discussed the utility of graphic captions for sounds that are off screen, with many suggesting they would not add graphic captions for sounds that are already on screen as those are already visually available. Considering these variations of appropriateness, participants suggested that a system could provide guidance such as when to add graphic captions for a sound, why someone might want to add a graphic caption, and which visualizations to include.

Participants talked about how it would be useful to break down the timestamps by scene for sounds that linger over scene changes when employing graphic captions, as their appropriateness may vary by scene. P5 suggested that it would be useful if the system could automatically detect scene changes and suggest time stamps for those scenes within a specific sound event. P4 also suggested that the system could automatically identify visual objects in the video to pin graphic captions to so that if the object moves, the graphic caption moves along with it.

7.2.3 Graphic captions may be beneficial, but also distracting. Many participants envisioned benefits from graphic captions, such as being able to add humor to entertainment or social media videos. For example, P7 said: "*The social media crowd, you know people who are using Facebook, TikTok and Instagram, will definitely get a kick out of this.*" However, some participants were also wary of the potential for visuals to "take away" or distract from the main content. For example, P10 noted: "*If I were to add any visuals, it would take away from the actual subject matter of the video*".

7.3 Automatic Systems Should Identify Important Sounds

When asked about what participants wanted from an automatic system that could identify non-speech sounds for them, most indicated they would want a system to identify *important* sounds. Some even highlighted the prototype as doing "a good job" of filtering important sounds when asked if they noticed any missing sounds in the results. P3, for instance, said: "*The car engine humming in the background, that was missing but I think that's good.*" Participants' inclinations to only include important sounds were also evident in their lack of interactions with the star function (a function described in Section 6.1). Participants found the purpose of the star unclear as they had only included important sounds already, and thus marking them with a star seemed redundant. P4, for example, said "*when you were explaining the starring feature in the tutorial earlier, I thought 'I don't know if that's something I would use' because I would only include sounds that are important.*" When discussing what constitutes an important sound, participants indicated sounds that are "noticeable," that "stood out," or that provided context for spoken content.

Many participants were not sure about whether "background" or "insignificant" sounds should also be included. However, P2 acknowledged the importance of including all sounds for someone who may not have the "privilege" to select which sounds to pay attention to. Thus, after engaging with the prototype, P2 decided to include all of the sounds in a video and star the important ones. Notably, when talking about what to do with "sounds that do not convey information" (P9), participants drew comparisons to alternative text (i.e. text added to images for screen readers to read). Both P4 and P9 talked about how guidelines for alternative text suggest marking images that do not convey information as "decorative." Thus, both participants considered sounds that are not important to be akin to decorative images.

Finally, many participants also talked about what information about a sound they would like an automatic system to identify. Participants mentioned wanting only *general* descriptions of what the sound may be as too much specificity would be likely to introduce errors. On the other hand, some participants highlighted the importance of obtaining accurate time stamps of the sounds from the system, which we explore in more detail in the next subsection. Most participants, however, indicated wanting both the descriptions and the timestamps for the sounds.

7.4 Dealing with Errors from Automatic Systems

In the conditions using the automatic system, some scenarios included both descriptions and timing errors. In some cases, participants suggested that the automatic system ended up being unhelpful when containing errors, with some deleting all the suggestions provided and starting from scratch instead. For instance, P10 commented: "*I could argue that [the automatic system] actually made things a little bit slower than just putting that things in manually.*" However, participants' discussions of errors suggest that their expectations differed when talking about the labeling of the sound versus the time stamps for the sounds.

7.4.1 The accuracy of timestamps may be the most important. Many participants highlighted accurate timestamps as an important feature of an automatic system. For example, P5 said "*Having the timings already sorted for you beforehand is very convenient. It takes out one of the biggest time-consuming parts of a process.*" Thus, participants seemed more sensitive to timing errors than errors in describing the sounds. For instance, P9 attributed the unhelpfulness of errors specifically to timestamps: "*Erroneous timestamps are not as useful.*" P1 also mentioned that timing errors would be time-consuming in longer videos: "*What if I had a 40-minute video? Am I gonna have to go through and look at every second?*" However, many participants agreed that having a "template" or "framework" with accurate timestamps is useful even when the descriptions of the sounds are incorrect.

7.4.2 Ambiguous sounds may be difficult to identify and describe, but strategies can help. Ambiguous sounds were at the core of participants' discussions of errors in the automatic labeling of the sounds. Many acknowledged that dealing with ambiguous sounds is difficult even for humans. For example, P2 said: "*I can see how the computer that's identifying the noise really has to be smart because [it is hard] even for me.*" For instance, one of the errors introduced by the automatic system was labeling an off-screen whining child as a "horse." Some participants actually trusted that this was a horse. P4 and P7 mentioned that even though they understood that a horse did not make sense in that context, the suggestion primed them and they could not think of the sound as something else. However, participants also discussed strategies to disambiguate certain sounds. For example, P8 found an error that labeled a trumpet as a mosquito understandable because visual information from the video was needed to disambiguate that sound: "*To tell the difference that this is a trumpet, you need a little bit of visual along with the sound to know.*" Others also talked about using spoken content in the video to disambiguate sounds. For example, P2 talked about correctly identifying that the sound labeled as a "horse" was a child using the spoken content from the video, which referred to a child. "*I guess it is confirmation that it is a baby,*" P2 said.

7.4.3 The ability to adjust the system's sensitivity may help in dealing with errors. When discussing errors from the automatic system, a few participants believed those errors were an issue of sensitivity. P6, for example, suggested that the errors were caused because the system may have been "too sensitive." Thus, these participants suggested adding a "reanalyze button" (P1) or a "sensitivity slider" (P7) that could help to reduce the errors from the system. P7 defined sensitivity in terms of the number of sounds identified, but also in terms of how "obvious" the sounds that the system identified are.

8 DISCUSSION AND TAKEAWAYS

This section summarizes the takeaways from our studies for those interested in captioning or visualizing non-speech sounds in user-generated content, including researchers investigating automatic captions. Considering that our findings include the perspectives of DHH viewers on new approaches to visualize non-speech sounds (i.e. graphic captions), some of our takeaways may also be insightful for industry professionals interested in visualizing non-speech sounds in professionally-produced content.

Be selective about non-speech sounds. The results from both of our studies suggest that both DHH and hearing users valued being selective about which non-speech sounds to include. DHH participants expressed interest in only having sounds that are *important*, as including every single detail may provide too much information. Hearing creators, in turn, described *important* sounds as what they would want an automatic system to identify because captioning all sounds in a video may be too time-consuming. These findings align with guidelines for creating closed captions suggests including non-speech sounds, which suggest only doing so *when necessary* (e.g. the BBC's captioning guidelines[12]). Our work further encourages researchers working on automatic sound event detection to consider importance estimations when detecting non-speech sounds in user-generated videos, and we provide insights about what constitutes an important sound. For instance, hearing creators discussed the criteria of whether a sound affects the spoken content of a video. Finally, estimations of sound importance or other sound qualities (e.g. volume) could be used to adjust the "sensitivity" of automatic systems and narrow down their results.

Include details, but balance with potential for distraction. Our results suggest that DHH participants are interested in having detailed information about non-speech sounds (e.g. the source, source location and changes in sounds) in text-based or graphic captions, which aligns with prior work on sound visualization [18] and the use visual-tactile feedback for non-speech information in captions [24]. However, DHH participants in our study worried that too many details in text-based captions, and graphic captions in general (which naturally included more details), could be distracting. Thus, those interested in captioning or visualizing non-speech sounds in their videos should consider including details, but be mindful of the potential for distracting viewers from the main content. A possible design intervention to support creators would be providing guidance within the user interface on how to structure the text-based captions (e.g. how many words to use, and which

[12] https://bbc.github.io/subtitle-guidelines/#Intonation-and-emotion

types of words) and how much detail to include.

Consider properties about the video, sound, and audience for choosing between text-based and graphic captions, and their respective benefits. Our findings reveal different factors to consider when choosing text-based or graphic captions. First, the type of video appeared to determine the appropriateness of graphic captions. The results from both studies suggested that graphic captions may be more appropriate for entertainment videos, while text-based captions may be more appropriate for "serious" videos. Future work can further explore whether varying the design of graphic captions to better match the content would affect viewers' and creators' preferences for the use of graphic captions. Other factors, such as the visibility of a sound on screen and demographic factors of the viewers (e.g. their age and hearing ability) are important too. While future work could explore each of these factors in more detail, our findings provide guidance for designers of these technologies to consider these factors about the videos, the sounds and their audiences. Our findings also shed light on what text-based and graphic captions may signify. The latter may serve to illustrate recognizable sounds and indicate the location of the sources, the precise timing of the sounds, as well as changes in the sounds. Text-based captions, on the other hand, may better describe harder-to-visualize abstract sounds where a greater level of verbal detail may be required.

Graphic captions should be optional and may need standardization. DHH participants suggested that graphic captions should be optional for viewers instead of being embedded in videos, an approach analogous to close captioning (as opposed to open captions, which are embedded in videos). Furthermore, DHH participants also suggested that graphic captions may be difficult to distinguish from other graphics in online videos. The ability to overlay them on demand may also support viewers in distinguishing which graphics are part of the video as opposed to graphic captions. However, current captioning technologies only support text (or animated text). Thus, new captioning or media formats, such as BBC's proposal for *Object-Based Media* [2, 3, 19], may need to be developed to support the addition and standardization of optional graphic captions.

Text-based captions of non-speech sounds may help distinguish uncaptioned videos. DHH participants also suggested that using text-based captions for non-speech sounds can support viewers to distinguish videos without spoken content from uncaptioned videos, as there are times when DHH viewers cannot tell if a video is not captioned or if it simply does not contain spoken content. If a video only contains non-speech sounds, and those sounds are captioned, viewers may conclude that the video does not contain spoken content (although there is still a possibility that a creator only captions non-speech sounds in a video with spoken content).

It is also possible that a video may contain unimportant (or decorative) non-speech sounds. Hearing creators' comparisons with alt-text, and the respective guidelines for decorative images that do not provide any information[13], suggest providing indications when videos only contain decorative sounds may help reduce the ambiguity of uncaptioned videos.

Providing accurate timestamps to creators is important, but general descriptions are helpful, too. Our results suggest that accurate time stamps are important when using an automatic system, as our hearing creators suggested that identifying these is the most time-consuming aspect of captioning in general. While descriptions were also important to many of our participants, general descriptions seemed more useful than specific ones given that specificity may introduce errors, especially when dealing with ambiguous sounds. Our results also suggest ways in which ambiguity can be reduced in automatic systems, which was a source of difficulty highlighted by YouTube for including non-speech sounds in their automatic captions [12]. More specifically, participants highlighted the use of semantic information from both visual and spoken content to reduce ambiguity, which aligns with current trends in multi-modal analysis and understanding.

9 CONCLUSION, LIMITATIONS, AND FUTURE WORK

In this paper, we presented two studies with DHH and hearing participants to explore their perspectives on captioning non-speech sounds, using text-based or graphic captions, in user-generated videos. Our findings include DHH participants' interests in having *important* non-speech sounds in these videos, while hearing creators also indicated an inclination toward only including *important* sounds when captioning non-speech sounds. Our findings also include trade-offs between text-based and graphic captions for captioning non-speech sounds, and potential factors for determining their appropriateness. Finally, we explored the use of automatic tools to support hearing creators when captioning non-speech sounds and identified guidance for future work in this area.

There were several limitations in our work, and avenues for future work. First, our work was qualitative, with a small sample size and a small selection of videos. We identified *potential* factors at play when determining the appropriateness of graphic captions (including the content and sound type, the sound location, viewers' demographic factors, and the graphic captions' style). However, future work should explore these factors using a larger selection of carefully controlled videos among a larger sample size of viewers.

Similarly, in our prototype study, participants were using sample videos provided by us. A study with content creators editing their own videos could reveal further insights relevant to real use cases. Our participants also had diverse levels of skills and experience. Future work investigating the preferences of participants with specific levels of skills or experience may yield further insights.

Our formative study suggested that automatic sound event detection may be helpful for DHH creators to caption non-speech sounds in their videos, which may introduce different challenges. Thus, future work should also explore its use among DHH creators and its implications.

ACKNOWLEDGMENTS

We thank Justin Salamon for helping with sound event detection and the Center for Accessibility and Inclusion Research (CAIR) at RIT for feedback on a early draft of the paper. This material is based

[13] https://webaim.org/techniques/alttext/#decorative

upon work supported by the National Science Foundation under awards No. 1822747 and 2125362.

REFERENCES

[1] Mike Armstrong, Andy Brown, Michael Crabb, Chris J Hughes, Rhianne Jones, and James Sandford. 2016. Understanding the diverse needs of subtitle users in a rapidly evolving media landscape. *SMPTE Motion Imaging Journal* 125, 9 (2016), 33–41.

[2] Mike Armstrong and Michael Crabb. 2017. Exploring ways of meeting a wider range of access needs through object-based media-workshop. In *Conference on Accessibility in Film, Television and Interactive Media, York, UK*.

[3] BBC Research & Development. [n.d.]. Object-Based Media. https://www.bbc.co.uk/rd/object-based-media. Accessed: 2022-06-01.

[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* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312921

[5] Larwan Berke, Sushant Kafle, and Matt Huenerfauth. 2018. *Methods for Evaluation of Imperfect Captioning Tools by Deaf or Hard-of-Hearing Users at Different Reading Literacy Levels.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173665

[6] Larwan Berke, Matthew Seita, and Matt Huenerfauth. 2020. Deaf and Hard-of-Hearing Users' Prioritization of Genres of Online Video Content Requiring Accurate Captions. In *Proceedings of the 17th International Web for All Conference* (Taipei, Taiwan) *(W4A '20)*. Association for Computing Machinery, New York, NY, USA, Article 3, 12 pages. https://doi.org/10.1145/3371300.3383337

[7] Danielle Bragg, Nicholas Huynh, and Richard E. Ladner. 2016. A Personalizable Mobile Sound Detector App Design for Deaf and Hard-of-Hearing Users. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 3–13. https://doi.org/10.1145/2982142.2982171

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

[9] Virginia Braun and Victoria Clarke. 2022. Thematic Analysis. https://www.thematicanalysis.net. Accessed: 2022-03-01.

[10] Andy Brown, Rhia Jones, Mike Crabb, James Sandford, Matthew Brooks, Mike Armstrong, and Caroline Jay. 2015. Dynamic Subtitles: The User Experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video* (Brussels, Belgium) *(TVX '15)*. Association for Computing Machinery, New York, NY, USA, 103–112. https://doi.org/10.1145/2745197.2745204

[11] Andy Brown, Jayson Turner, Jake Patterson, Anastasia Schmitz, Mike Armstrong, and Maxine Glancy. 2017. Subtitles in 360-Degree Video. In *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video* (Hilversum, The Netherlands) *(TVX '17 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 3–8. https://doi.org/10.1145/3084289.3089915

[12] Sourish Chaudhuri. 2017. Adding sound Effect information to Youtube captions. https://ai.googleblog.com/2017/03/adding-sound-effect-information-to.html Accessed: 2021-06-01.

[13] Karen Collins and Peter J. Taillon. 2012. Visualized sound effect icons for improved multimedia accessibility: A pilot study. *Entertainment Computing* 3, 1 (2012), 11–17. https://doi.org/10.1016/j.entcom.2011.09.002

[14] Michael Crabb, Rhianne Jones, and Mike Armstrong. 2015. The Development of a Framework for Understanding the UX of Subtitles. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery, New York, NY, USA, 347–348. https://doi.org/10.1145/2700648.2811372

[15] Michael Crabb, Rhianne Jones, Mike Armstrong, and Chris J. Hughes. 2015. Online News Videos: The UX of Subtitle Position. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery, New York, NY, USA, 215–222. https://doi.org/10.1145/2700648.2809866

[16] Sofia Enamorado. 2018. CVAA & FCC closed Captioning requirements for online video. https://www.3playmedia.com/blog/final-cvaa-and-fcc-online-video-closed-captioning-rules/ Accessed: 2021-06-01.

[17] Deborah I Fels, Daniel G Lee, Carmen Branje, and Matthew Hornburg. 2005. *Emotive Captioning and Access to Television.* https://doi.org/10.1145/3173574.3173665

[18] 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.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300276

[19] Benjamin M. Gorman, Michael Crabb, and Michael Armstrong. 2021. Adaptive Subtitles: Preferences and Trade-Offs in Real-Time Media Adaption. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 733, 11 pages. https://doi.org/10.1145/3411764.3445509

[20] Michael Gower, Brent Shiver, Charu Pandhi, and Shari Trewin. 2018. Leveraging Pauses to Improve Video Captions. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 414–416. https://doi.org/10.1145/3234695.3241023

[21] Dhruv Jain, Sasa Junuzovic, Eyal Ofek, Mike Sinclair, John Porter, Chris Yoon, Swetha Machanavajhala, and Meredith Ringel Morris. 2021. A Taxonomy of Sounds in Virtual Reality. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 160–170. https://doi.org/10.1145/3461778.3462106

[22] Dhruv Jain, Angela Lin, Rose Guttman, Marcus Amalachandran, Aileen Zeng, Leah Findlater, and Jon Froehlich. 2019. *Exploring Sound Awareness in the Home for People Who Are Deaf or Hard of Hearing.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300324

[23] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894. https://doi.org/10.1109/TASLP.2020.3030497

[24] Raja S. Kushalnagar, Gary W. Behm, Joseph S. Stanislow, and Vasu Gupta. 2014. Enhancing Caption Accessibility through Simultaneous Multimodal Information: Visual-Tactile Captions. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility* (Rochester, New York, USA) *(ASSETS '14)*. Association for Computing Machinery, New York, NY, USA, 185–192. https://doi.org/10.1145/2661334.2661381

[25] Daniel G. Lee, Deborah I. Fels, and John Patrick Udo. 2007. Emotive Captioning. *Comput. Entertain.* 5, 2, Article 11 (April 2007), 15 pages. https://doi.org/10.1145/1279540.1279551

[26] Tara Matthews, Janette Fong, F. Wai-Ling Ho-Ching, and Jennifer Mankoff. 2006. Evaluating non-speech sound visualizations for the deaf. *Behaviour & Information Technology* 25, 4 (2006), 333–351. https://doi.org/10.1080/01449290600636488 arXiv:https://doi.org/10.1080/01449290600636488

[27] John McGowan, Grégory Leplâtre, and Iain McGregor. 2017. CymaSense: A Real-Time 3D Cymatics-Based Sound Visualisation Tool. In *Proceedings of the 2017 ACM Conference Companion Publication on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 270–274. https://doi.org/10.1145/3064857.3079159

[28] Carol Padden and Tom Humphries. 2005. *Inside Deaf Culture.* Harvard University Press. http://www.jstor.org/stable/j.ctvjz83v3

[29] S. J. Parault and H. M. Williams. 2010. Reading Motivation, Reading Amount, and Text Comprehension in Deaf and Hearing Adults. *Journal of Deaf Studies and Deaf Education* 15, 2 (2010), 120–135. https://doi.org/10.1093/deafed/enp031

[30] C. B. Traxler. 2000. The Stanford Achievement Test, 9th Edition: National Norming and Performance Standards for Deaf and Hard-of-Hearing Students. *Journal of Deaf Studies and Deaf Education* 5, 4 (Jan 2000), 337–348. https://doi.org/10.1093/deafed/5.4.337

[31] M. Wald. 2011. *Crowdsourcing Correction of Speech Recognition Captioning Errors.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1969289.1969318

[32] Dawn Walton, Georgianna Borgna, Marc Marschark, Kathryn Crowe, and Jessica Trussell. 2019. I am not unskilled and unaware: deaf and hearing learners' self-assessments of linguistic and nonlinguistic skills. *European Journal of Special Needs Education* 34, 1 (2019), 20–34. https://doi.org/10.1080/08856257.2018.1435010 arXiv:https://doi.org/10.1080/08856257.2018.1435010

[33] Fangzhou Wang, Hidehisa Nagano, Kunio Kashino, and Takeo Igarashi. 2015. Visualizing video sounds with sound word animation. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME.2015.7177422

[34] Noah Wang. 2017. Visualizing sound effects. https://youtube-eng.googleblog.com/2017/03/visualizing-sound-effects.html Accessed: 2021-06-01.

[35] Sean Zdenek. 2015. *Reading sounds: Closed-captioned media and popular culture.* University of Chicago Press.