When to Add Human Narration to Photo-Sharing Social Media

Conference Paper · October 2020

DOI: 10.1145/3373625.3418013

Lawrence Kim
Simon Fraser University
36 PUBLICATIONS **826** CITATIONS
SEE PROFILE

Abena Boadi-Agyemang
Carnegie Mellon University
6 PUBLICATIONS **44** CITATIONS
SEE PROFILE

Alexa Siu
Stanford University
21 PUBLICATIONS **831** CITATIONS
SEE PROFILE

When to Add Human Narration to Photo-Sharing Social Media

LAWRENCE H. KIM[*], Stanford University
ABENA BOADI-AGYEMANG[*], Stanford University
ALEXA F. SIU, Stanford University
JOHN TANG, Stanford University and Microsoft Research

Social media platforms facilitate communication through sharing photos and videos. The abundance of visual content creates accessibility issues, particularly for people who are blind or have low vision. While assistive technologies like screen readers can help when alt-text for images is provided, synthesized voices lack the human element that is important for social interaction. Here, we investigate when it makes the most sense to use human narration as opposed to a screen reader to describe photos in a social media context. We explore the effects of voice familiarity (i.e., whether you hear the voice of someone you know) and the perspective of the description (i.e., first vs. third person point-of-view (POV)). Preliminary study suggests that users prefer hearing from a person they know when the content is described in first person POV, whereas synthesized voice is preferred for content described in third person POV.

Additional Key Words and Phrases: Accessible Social Media, Human Narration

ACM Reference Format:
Lawrence H. Kim, Abena Boadi-Agyemang, Alexa F. Siu, and John Tang. 2020. When to Add Human Narration to Photo-Sharing Social Media. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20), October 26–28, 2020, Virtual Event, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3373625.3418013

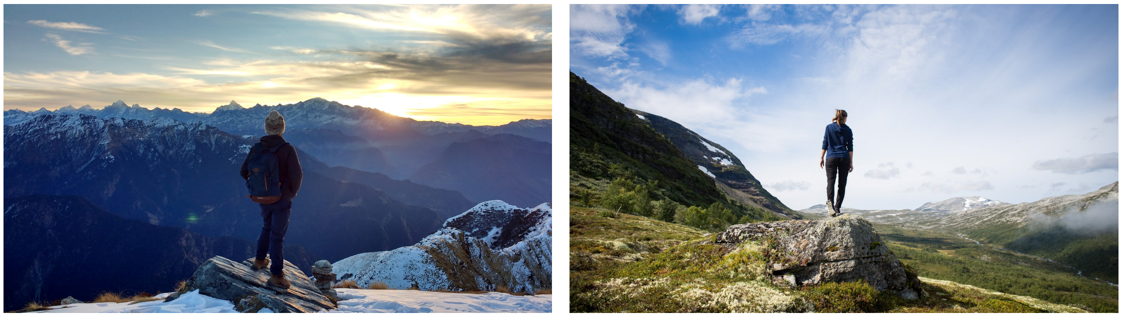
1 INTRODUCTION

Social media is a popular forum for contemporary communication. Photo-sharing platforms, such as Instagram, however pose accessibility challenges for users, especially those who are blind or low vision (BLV) [2, 6, 10]. Many BLV users rely on assistive technologies such as a screen reader to navigate and interpret the visual content on social media [3]. Screen readers leverage alternative text (alt-text) embedded in images to provide audio descriptions for BLV users.

The lack of alt-text-embedded images prevents BLV users from engaging with the primarily visual content of social media. Previous work has shown that some of the reasons for the lack of alt-text are that users may forget to add alternative text, do not have time to add it, or do not know what to include when writing these descriptions [7]. In 2018, Instagram enabled users to create custom alt-text and launched default automatic alt-text (AAT) on images for the benefit of BLV users [1]. While these efforts can improve the accessibility of photo-based social media, AAT, which relies on computer vision technology, is often vague and imprecise [11]. Therefore, human-authored alt-text still has the potential to provide rich and contextually appropriate descriptions of images for BLV users.

Social media accessibility can be augmented with text-based alternatives, such as audio. In June 2020, Twitter launched audio tweets, which allow users to share custom voice recordings [9]. Despite the issues with audio tweeting

[*] Both authors contributed equally to this research.



(a) Description in *first person perspective*: ”I stand atop a tilted rock surrounded by ice. My back is turned as I overlook the sunset beyond the mountain range in the distance.” © Gaurav K

(b) Description in *third person perspective*: ”A woman stands atop a rock surrounded by greenery. Her back is turned as she gazes at the sprawling plain.” © Philipp Lublasser

Fig. 1. Two comparable images used for the study that fall under the activities category. The descriptions are similar but are written in different perspectives.

(e.g., being inaccessible to users who are deaf or hard-of-hearing), this feature could be used to create succinct audio descriptions of photos. Human-authored audio descriptions of images can be directly consumed by BLV users without the need for automatic or manual creation of alt-text or the interpretation of a screen reader. Furthermore, prior work suggests that human-authored narrations are preferred over the synthesized voice of a screen reader, even for regular screen reader users [4, 5].

We explore the conditions for using human narration for photos on social media. We study the effects of familiarity with the content author (e.g., familiar or unfamiliar) and perspective of the descriptions (i.e., first person vs. third person) on preferred voice (i.e., synthesized vs. human). From our preliminary study, we find that people may prefer human narration when it is described in a first person POV but prefer screen reader voice when narrated in third person POV.

2 EVALUATION

Prior work has shown that people prefer human voice audio description over text read by a screen reader for both Instagram [4] and films [5]. However, social media encompasses many different types of content, ranging from personal reflections to news reports. In addition, human audio description requires additional input from users which may not always be available. Thus, we need to understand when it makes the most sense to add human narration to photo-based social media.

2.1 Independent Variables: Voice Type and Perspective

Social media differs from other forms of media in that users may consume information from not only strangers but also from people they know, including close friends and family members. We hypothesize that the element of familiarity will influence whether consumers prefer a human narration or a description from a screen reader. Thus, we present and examine three *voice types*: familiar human voice, unfamiliar human voice, and synthesized voice.

Since the purpose of alt-text is to provide an objective description of the contents of an image, they are most often written in the third person POV. However, in the context of social media, first person description may be more appropriate, especially for personal content. Thus, we would like to understand how users perceive different *perspectives* (i.e., first person vs. third person descriptions) and whether the perspective affects their preferred voice type.

2.2 Content Preparation and Procedure

Instagram photos fall into roughly eight categories [8]. We focus on the three most popular categories (selfies, friends, and activities) which all center around people. We used a set of two comparable photos for each category, one of which is shown in Fig. 1. The descriptions were also written similarly but in a different perspective for each set.

In order to provide narration from a familiar person who is not a well-known figure such as Barack Obama, we recruited people who are BLV and know one of the authors. We recorded and used that author's voice for the familiar voice type condition. To control for other factors, we recruited another person of the same gender for the unfamiliar voice type condition, and used a synthesized voice of the same perceived gender. To reduce potential bias, authors who did not provide the narrations conducted the study.

We developed an interface with a similar layout to Instagram. Participants could navigate different images and play associated audio using their keyboard. We conducted the study over a video call to share the prototype and receive feedback on the user experience. In total, there were 3 (image types) x 3 (voice type) x 2 (perspective) = 18 conditions. We presented the corresponding narration to the participants in a counter-balanced order across the different image types. After presenting all stimuli, we gathered qualitative responses about their experience through an online interview and a post-study survey. Participants ranked their preferred voice type from 1 to 3 (1 being the most preferred and 3 being the least) based on how close they felt to the content author, how pleasant the description sounded, and how much they enjoyed the overall experience and provided explanations to support their preferences.

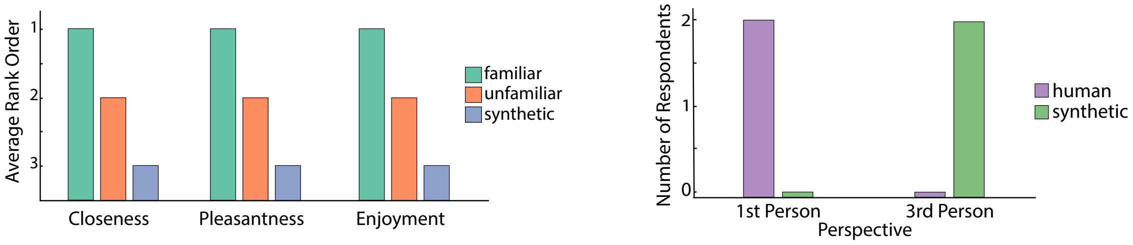
3 PRELIMINARY RESULTS & DISCUSSION

We recruited two participants who are blind or have low vision. Both participants are regular screen reader users and use social media infrequently because of a lack of accessibility. After they experienced all the conditions through our interface, we asked them about their thoughts on the presented voice types (i.e., familiar human voice, unfamiliar human voice, and synthesized voice) and perspectives (i.e., first person vs. third person descriptions).

3.1 Perception of *Voice Types*

The participants expressed a preference for the familiar voice followed by the unfamiliar voice during the interview and the post-study survey, as shown in Fig. 2a. P1 noted that *"It was a little easier to make the connection with the content of the post when the narrating voice was someone I knew"*, while P2 said *"When I hear [a person I know] ... I imagine [them] doing it [an activity]"*. While the familiar voice ranked the highest, the degree of familiarity with the voice of the content author may influence the participants' sense of social connection with the author as well as their perception of the image description. P1 explained that *"If one of my closer friends had been narrating, the connection would have been even stronger."* On the other hand, P2 thought all the narrations were describing someone else's photo, and thus *"preferred hearing it from a person I don't know"* to avoid forming a biased mental image that would occur if a person that P2 knows was narrating. For future studies, it would be worth testing various degrees of familiarity as both participants in our study were only acquaintances of the content author of the familiar voice condition, not close friends.

The synthesized voice received the lowest rankings for closeness, pleasantness and enjoyment, as shown in Fig. 2a. P1 expressed that *"If the screen-reading voice had been closer to what I use on a daily basis ... [in terms of] rate, pitch, [or] voice profile ... I think I would have found it more pleasant."* P2 was more definitive in their stance: *"I always use [a] screen reader so I'm used to it ... I always prefer a human voice over the synthesized voice."* Frequent use of a screen reader may account for consistently low rankings of the synthesized voice across conditions since this voice type lacks novelty.



(a) Rank of voice type for closeness, pleasantness, and enjoyment

(b) Preferred voice type for each perspective

Fig. 2. (a) For all three measures, both users ranked familiar voice the highest followed by unfamiliar and synthetic voices respectively. (b) For first person perspective, both users preferred human voice whereas synthetic voice was preferred for third person perspective.

Additionally, the non-customized synthesized voice may have resulted in a lower ranking because it is not what the participants are used to hearing.

3.2 Perception of *Perspectives* and Interaction with *Voice Type*

The participants shared the factors that influenced their preferences concerning first person and third person descriptions. P1 expressed that *"It was easier for me to get a mental image when the description was in third person...almost like listening to an audio book and having the narrator construct an image in my mind with each detail."* P1 elaborated that *"First person narration made me feel almost like I was in a conversation with the person speaking."* Additionally, P1 prefers *"[the] first person if the speaker was describing themselves in a post, [and] third person description if someone I knew wasn't in the picture [or] post being described."* Perspective use depends on the context. First person descriptions are useful for establishing a connection between viewer and content creator. Third person descriptions are helpful in creating unbiased mental images or when a content creator is describing a photo featuring someone else.

Furthermore, the participants shared that they prefer first person descriptions be in the human voice and third person in a synthesized voice as shown in Fig. 2b. P1 stated that *"I am more used to screen-readers reading in the third person – [such as] alt-text for images"*. For first person descriptions, P2 said that *"if it is a human voice, then I would prefer [first person], then [I] feel closer [to them]"*, while P1 preferred *"the familiar [human] voice for first person because it felt more like a natural interaction [or] conversation I was having with the narrator about their post."* This suggests, contrary to results from prior work [4, 5], that human voice is not always preferred over synthesized voice but depends on the context such as perspective.

4 CONCLUSION

Many social media platforms are visual, creating accessibility issues, particularly for people who are blind or visually impaired. While assistive technologies such as a screen reader help bridge the gap, the synthesized voice is often unsuitable for social interaction. To understand when to use human voice, we explore the effects of familiarity and the perspective of the description. Preliminary results suggest that users prefer hearing from people they know when the content is described in first person POV, but prefer synthesized voice for content in third person POV. In light of social media platforms like Twitter leveraging human narration, our preliminary findings concerning the preferences among BLV users suggest that social media posts that include a human-generated audio description in the first person POV might lead to a richer story-telling social media experience for all.

REFERENCES

[1] 2018. Improved Accessibility Through Alternative Text Support. https://about.instagram.com/blog/announcements/improved-accessibility-through-alternative-text-support

[2] Cynthia L Bennett, Jane E, Martez E Mott, Edward Cutrell, and Meredith Ringel Morris. 2018. How teens with visual impairments take, edit, and share photos on social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[3] Maria Claudia Buzzi, Marina Buzzi, Barbara Leporini, and Fahim Akhter. 2010. Is Facebook really" open" to all?. In *2010 IEEE International Symposium on Technology and Society*. IEEE, 327–336.

[4] João Marcelo dos Santos Marques, Luiz Fernando Gopi Valente, Simone Bacellar Leal Ferreira, Claudia Cappelli, and Luciana Salgado. 2017. Audio Description on Instagram: Evaluating and Comparing Two Ways of Describing Images for Visually Impaired.. In *ICEIS (3)*. 29–40.

[5] Anna Fernández-Torné and Anna Matamala. 2015. Text-to-speech vs. human voiced audio descriptions: a reception study in films dubbed into Catalan. *The Journal of Specialised Translation* 24 (2015), 61–88.

[6] Kristin Skeide Fuglerud, Ingvar Tjøstheim, Birkir Rúnar Gunnarsson, and Morten Tollefsen. 2012. Use of social media by people with visual impairments: usage levels, attitudes and barriers. In *International Conference on Computers for Handicapped Persons*. Springer, 565–572.

[7] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M Kitani, and Jeffrey P Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference*. 549–559.

[8] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What we instagram: A first analysis of instagram photo content and user types. In *Eighth International AAAI conference on weblogs and social media*.

[9] Maya Patterson and Rémy Bourgoin. 2020. Your Tweet, your voice. https://blog.twitter.com/en_us/topics/product/2020/your-tweet-your-voice.html

[10] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How blind people interact with visual content on social networking services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1584–1595.

[11] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1180–1192.