

Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

10-2021

Uncovering patterns in reviewers' feedback to scene description authors

Rosiana NATALIE

Singapore Management University, rosianan@smu.edu.sg

Jolene Kar Inn LOH

Singapore Management University, jolene.loh.2019@scis.smu.edu.sg

Huei Suen TAN

Singapore Management University, hstan.2017@sis.smu.edu.sg


Joshua Shi-hao TSENG


Singapore Management University, joshuatseng.2017@sis.smu.edu.sg

Hernisa KACORRI

University of Maryland at College Park

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 See next page for additional authors

 Part of the [Graphics and Human Computer Interfaces Commons](#), and the [Software Engineering Commons](#)

Citation

NATALIE, Rosiana; LOH, Jolene Kar Inn; TAN, Huei Suen; TSENG, Joshua Shi-hao; KACORRI, Hernisa; and HARA, Kotaro. Uncovering patterns in reviewers' feedback to scene description authors. (2021). *ASSETS '21: Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility, Virtual, October 18-22*. 1-4.

Available at: https://ink.library.smu.edu.sg/sis_research/8148

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Author

Rosiana NATALIE, Jolene Kar Inn LOH, Huei Suen TAN, Joshua Shi-hao TSENG, Hernisa KACORRI, and Kotaro HARA



Published in final edited form as:

ASSETS. 2021 ; 93: . doi:10.1145/3441852.3476550.

Uncovering Patterns in Reviewers' Feedback to Scene Description Authors

Rosiana Natalie,

Singapore Management University, Singapore

Jolene Loh Kar Inn,

Singapore Management University, Singapore

Tan Huei Suen,

Singapore Management University, Singapore

Joshua Tseng Shi Hao,

Singapore Management University, Singapore

Hernisa Kacorri,

University of Maryland, College Park, USA

Kotaro Hara

Singapore Management University, Singapore

Abstract

Audio descriptions (ADs) can increase access to videos for blind people. Researchers have explored different mechanisms for generating ADs, with some of the most recent studies involving paid novices; to improve the quality of their ADs, novices receive feedback from reviewers. However, reviewer feedback is not instantaneous. To explore the potential for real-time feedback through automation, in this paper, we analyze 1, 120 comments that 40 sighted novices received from a sighted or a blind reviewer. We find that feedback patterns tend to fall under four themes: (i) **Quality**; commenting on different AD quality variables, (ii) **Speech Act**; the utterance or speech action that the reviewers used, (iii) **Required Action**; the recommended action that the authors should do to improve the AD, and (iv) **Guidance**; the additional help that the reviewers gave to help the authors. We discuss which of these patterns could be automated within the review process as design implications for future AD collaborative authoring systems.

Keywords

Audio Description; visual impairment; video accessibility; collaborative writing

1 INTRODUCTION

Videos rely mainly on visual communication, which is typically inaccessible for blind people. Audio Description (AD) describes the inaudible visual events happening in the

scenes to blind people in the form of speech, but many videos that exist today lack ADs and remain inaccessible. For example, according to the American Council of the Blind, only about 2,760 out of 75 million videos (0.004%) on Amazon Prime Video come with ADs [10, 15]. Hiring professionals to generate high-quality ADs remains a gold standard for making video accessible, but it is costly and time-consuming (US\$ 12 to US\$75 per video minute [12, 18], with the turnaround time from days to weeks [18]).

Prior work explored ways to rely on more affordable and readily available novices to create ADs [1, 2, 4, 6, 7, 11, 16, 22–24]. For example, prior work by Kobayashi *et al.* [6] and Branje and Fels [1] showed that novices could be AD authors who can provide cost-effective ADs with a reasonable quality. More recently, researchers have examined ways to incorporate automated methods to reduce the burden of authoring ADs or fully automating the AD generation. [20, 23, 24].

In our prior work, we created ViScene, which enables mixed-ability collaboration between a pair of *author* and *reviewer* to collaboratively write scene descriptions (SD)—text that is transformed into audio through text-to-speech to become ADs [13, 14]. Using ViScene, an *author* wrote initial SDs, a *reviewer* then read the SDs and commented on how to improve them, and the author revised (Video Figure). We reported that the reviewer’s feedback helped authors improve SDs’ qualities like Descriptive, Objective, Referable and Clarity. We also showed our approach could reduce the cost and turnaround time of SD authoring compared to professional AD production, but it was evident that manual review limits the scalability of SDs creation.

As a step toward understanding whether we could automate the ViScene’s review process, we studied patterns in the comments that ViScene’s reviewers gave to SD authors. We conducted content analysis to analyze $N=1,120$ review feedback that two reviewers (one sighted and one blind reviewer) gave in the study described in our prior work [14]. Through iterative open coding and axial coding, we identified four themes in reviewers’ feedback. (i) Quality. The reviewers commented on nine qualities of SDs; they commented on some qualities more often than the other. (ii) Speech Act. The speech acts [5] of comments included either directive *instruction* or *questions*, or *warning* or acknowledgement *compliment*. (iii) Required Action. Directive feedback specified three types of the required actions: *revision*, *add information*, and *fix grammar*. (iv) Guidance. The reviewers often accompanied the Directive advice and questions by guiding examples, revision suggestions, or probing options. We will discuss how these findings could inform the design of future automated reviewing system to guide novices to generate SD.

2 METHOD

Data.

We used the review data from the prior research that studied ViScene [14] to uncover the patterns in how reviewers give feedback to novice SD authors. In the prior study, the authors wrote SDs for the following three videos and the reviewers gave comments on them: (i) an *Explainer* video about the importance of color contrast (7 scenes) [19], (ii) an *Instructional* video of a corner bookmark origami (12 scenes) [17], and (iii) an *Advertisement* video

that promotes Subaru car (9 scenes) [8]. We had 1, 120 scene-level SDs. We had the same number of reviews provided by two reviewers; a blind reviewer reviewed the half of the SDs and a sighted reviewer reviewed another half (*i.e.*, *Explainer*: (Blind, Sighted) = (140, 140) reviews, *Instructional*: (Blind, Sighted) = (240, 240) reviews, *Advertisement*: (Blind, Sighted) = (180, 180) reviews).

Analysis.—To understand the reviewers' feedback patterns, two researchers of this paper analyzed the reviews with content analysis. The two researchers first independently open coded a subset of the reviews (100 reviews; 50 from the sighted reviewer and 50 from the blind reviewer), then met to perform axial coding where they discussed to extract initial set of themes. Four themes emerged: *Quality*, *Speech Act*, *Required Actions*, and *Guidance*. To test if this set of themes is comprehensive, the same researchers performed another round of open and axial coding with additional 150 reviews. No new themes emerged in this round. In total, the two researchers individually coded 250 reviews ($0.81 \leq \kappa \leq 0.92$, $Mean = 0.865$, $SD = 0.045$;). The two researchers met, resolved all disagreements, and organized the emerged theme into a codebook. The two researchers coded additional 250 reviews to test the robustness of the codebook; this time, we observed a much higher agreement ($0.94 \leq \kappa \leq 0.99$, $Mean = 0.865$ ($SD = 0.022$)). Again, the disagreements were resolved through consensus and the codebook was updated for the final adjustment. Then, one researcher used the codebook to code the remaining reviews.

- Quality.** The reviewers commented on nine different SD quality dimensions. The quality dimensions were what we suggested the reviewers to comment on in the previous study [14]. Frequencies of the comments on the quality dimensions were: $N_{Descriptive} = 589$, $N_{Objective} = 78$, $N_{Succinct} = 9$, $N_{Learning} = 76$, $N_{Sufficient} = 470$, $N_{Interest} = 58$, $N_{Clarity} = 291$, $N_{Accurate} = 105$, and $N_{Referable} = 151$. Both reviewers most frequently commented on Descriptiveness ($N_{Blind} = 195$, $N_{Sighted} = 394$; Table 1). For example, the sighted reviewer asked, “What was the woman’s expression? What can you tell from her expression?” In contrast, only nine comments were about succinctness of SDs ($N_{Blind} = 8$, $N_{Sighted} = 1$). The result was expected because ViScene was equipped with succinctness visualization, guiding the authors to provide a more succinct SD [14].
- Speech Act.** All of the reviews’ speech acts fell under *directive* and *acknowledgement*, (*i.e.*, the categories explained in [5]). Directive reviews comprises of: (i) *Instruction* ($N = 902$), (ii) *Question* ($N = 77$), (iii) *Warning* ($N = 45$). The reviewers also acknowledged SDs in a form of *Compliment* ($N = 392$). One review could have multiple speech acts. The most common acts were *Instructions* on what the authors need to do ($N_{Blind} = 435$, $N_{Sighted} = 467$), followed by *Compliments* ($N_{Blind} = 204$, $N_{Sighted} = 188$), *Warnings* ($N_{Blind} = 26$, $N_{Sighted} = 19$), and *Questions* ($N_{Blind} = 13$, $N_{Sighted} = 64$). The sighted reviewer often instructed the authors to add more description about the scene (*e.g.*, “Describe more about the setting of the restaurant”). The blind reviewer suggested the authors to be more mindful about punctuation to make SDs sounds more natural when they were synthesized by text-to-speech. Also, the reviewers

complimented the authors when they described the scenes well ($N_{Blind} = 205$, $N_{Sighted} = 188$), like “No comment about this scene description. It is clear and easy to understand.” - *Blind*

- **Required Action.** The directive speech act specified at least one of three types of required actions: (i) *Revise* ($N = 352$), (ii) *Add Information* ($N = 743$), and (iii) *Fix Grammar* ($N = 56$). For both blind and sighted reviewers, the most common requests were to ask the authors to add more information ($N_{Blind} = 298$, $N_{Sighted} = 445$). We observed the blind reviewer requested more revisions ($N_{Blind} = 269$) compared to the sighted reviewer ($N_{Sighted} = 83$).
- **Guidance.** We observed that the reviewers occasionally used three ways to guide authors to rectify the problems: *Suggestion* ($N = 920$), *Example* ($N = 552$) and *Probing Option* to nudge authors to clarify contents of SDs ($N = 20$). To guide the authors in revising SDs, the reviewers most commonly gave Suggestions ($N_{Blind} = 447$, $N_{Sighted} = 473$), followed by Example ($N_{Blind} = 308$, $N_{Sighted} = 244$), and probing options ($N_{Blind} = 4$, $N_{Sighted} = 29$).

We noticed the discrepancy between what quality dimensions the blind and sighted reviewers commented on. The sighted reviewer mostly commented on the descriptiveness ($N_{Blind} = 195$, $N_{Sighted} = 394$), while the blind reviewer mostly commented on SD’s sufficiency ($N_{Blind} = 291$, $N_{Sighted} = 179$). The blind reviewer also often commented on the clarity of the SDs ($N_{Blind} = 267$, $N_{Sighted} = 24$). While the sighted reviewer gave zero feedback on the Interest quality, the blind reviewer gave 58 feedback on it.

We also noticed a difference in the actions the blind and sighted reviewers requested from the authors. While both reviewers requested the authors to add more information, the sighted reviewer wrote this type of feedback 1.5 times more frequently than to the blind reviewer ($N_{Blind} = 298$, $N_{Sighted} = 445$). Instead, the blind reviewer gave feedback to revise the SDs 3.2 times more often than the sighted reviewer ($N_{Blind} = 269$, $N_{Sighted} = 83$). Lastly, the blind reviewer gave 56 comments on fixing grammar but sighted reviewer gave none.

3 CONCLUSION AND DESIGN IMPLICATIONS

We summarized four themes in the reviewers’ comments to the SD authors, which are Quality, Speech Act, Required Action, and Guidance. We noticed a relationship between Quality and Required Action. Requests to improve the Descriptive and the Sufficient qualities were dominant, and the closer look at the data showed that the reviewers suggested the authors to improve these qualities by adding more information to SDs. We also found that there is a large difference in the number of *Revise* sub-theme between blind and sighted reviewers ($N = 186$). The large difference was mostly because the blind reviewer asked the authors to revise the sentence’s phrasing to improve Clarity ($N = 243$). Lastly, the blind reviewer requested to fix grammar more frequently than the sighted reviewer ($N_{Blind} = 56$, $N_{Sighted} = 0$). It was likely related to the Interest sub-theme; we observed that the incorrect grammar affected the enjoyment aspects of the video.

We believe our result informs the design of the future system to generate automatically reviews. For example, *Descriptiveness*, *Clarity*, and *Sufficiency* were the most common

quality feedback. The result suggests that we could significantly reduce the review workload if we could automatically create high-quality reviews on these quality dimensions. Future work could investigate how to incorporate off-the-shelf computer vision (*e.g.*, [9, 21]) and natural language processing algorithms (*e.g.*, [3]) to generate reviews on these qualities automatically. The system could then instruct and guide SD authors on what actions they should take (*e.g.*, when an SD is not descriptive enough, an auto-generated comment could ask the author to add more information).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore under its Strategic Capabilities Research Centres Funding Initiative, Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant, and a Lee Kong Chian Fellowship. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. Hernisa Kacorri is partially supported by the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR), ACL, HHS (#90REGE0008).

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

REFERENCES

- [1]. Branje Carmen J and Fels Deborah I. 2012. Livedescribe: can amateur describers create high-quality audio description? *Journal of Visual Impairment & Blindness* 106, 3 (2012), 154–165.
- [2]. Campos Virginia P, de Araújo Tiago MU, de Souza Filho Guido L, and Gonçalves Luiz MG. 2020. CineAD: a system for automated audio description script generation for the visually impaired. *Universal Access in the Information Society* 19, 1 (2020), 99–111.
- [3]. Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.
- [4]. GBH. 2020. CADET - Caption and Descriptive Editing Tool. <https://www.wgbh.org/foundation/what-we-do/ncam/cadet>. Accessed: 2020-11-6.
- [5]. Jurafsky Dan and Martin James H.. 2019. *Speech & language processing*. (2019). Draft edition, Refer to: <https://web.stanford.edu/~jurafsky/slp3/>.
- [6]. Kobayashi Masatomo, Fukuda Kentarou, Takagi Hironobu, and Asakawa Chieko. 2009. Providing synthesized audio description for online videos. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. 249–250.
- [7]. Kobayashi Masatomo, O’Connell Trisha, Gould Bryan, Takagi Hironobu, and Asakawa Chieko. 2010. Are synthesized video descriptions acceptable?. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*. 163–170.
- [8]. Lasher Lew. 2018. Subaru commercial | See the World | with audio description. <https://www.youtube.com/watch?v=flu6u988kh0> Accessed: 2021-06-15.
- [9]. Lei Jie, Wang Liwei, Shen Yelong, Yu Dong, Berg Tamara, and Bansal Mohit. 2020. MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2603–2614.
- [10]. Market.US. 2020. Amazon Prime Video Statistics and Facts. <https://market.us/statistics/online-video-and-streaming-sites/amazon-prime-video/>. Accessed:2021-06-15.

- [11]. 3Play Media. 2020. 3Play Plugin. <https://www.3playmedia.com/services/features/plugins/3play-plugin/>. Accessed: 2020-11-6.
- [12]. Mikul Chris. 2010. Audio description background paper. Media Access Australia (2010).
- [13]. Natalie Rosiana, Jarjue Ebrima, Kacorri Hernisa, and Hara Kotaro. 2020. ViScene: A Collaborative Authoring Tool for Scene Descriptions in Videos. In The 22nd International ACM SIGACCESS Conference on Computers and Accessibility. 1–4.
- [14]. Natalie Rosiana, Loh Jolene, Tan Huei Suen, Tseng Joshua, Chan Ian Luke Yi-ren, Jarjue Ebrima, Kacorri Hernisa, and Hara Kotaro. 2021. Efficacy of Collaborative Authoring of Video Scene Descriptions. To appear in The 23rd International ACM SIGACCESS Conference on Computers and Accessibility.
- [15]. American Council of The Blind. 2021. Amazon Prime Video Audio Described Titles. <https://acb.org/adp/amazonad.html>. Accessed:2021-06-15.
- [16]. Pavel Amy, Reyes Gabriel, and Bigham Jeffrey P. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. 747–759.
- [17]. Easy Peasy and Fun. 2017. How to Make DIY Origami Corner Bookmarks. <https://www.youtube.com/watch?v=hO4J1GjPQFw> Accessed: 2021-06-15.
- [18]. Thompson Terril. 2017. My Audio Description Talk @ CSUN). <http://terillthompson.com/813>. Accessed: 2020-11-6.
- [19]. W3C Web Accessibility Initiative (WAI). 2016. Web Accessibility Perspectives: Colors with Good Contrast - Audio Described Version. <https://www.youtube.com/watch?v=a9kNUv6N8Rk> Accessed: 2021-06-15.
- [20]. Wang Yujia, Liang Wei, Huang Haikun, Zhang Yongqi, Li Dingzeyu, and Yu Lap-Fai. 2021. Toward Automatic Audio Description Generation for Accessible Videos. (2021).
- [21]. Xu Kelvin, Ba Jimmy, Kiros Ryan, Cho Kyunghyun, Courville Aaron, Salakhudinov Ruslan, Zemel Rich, and Bengio Yoshua. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning. PMLR, 2048–2057.
- [22]. YouDescribe. 2020. YouDescribe. <https://youdescribe.org/support/tutorial>. Accessed: 2020-11-6.
- [23]. Yuksel Beste F, Fazli Pooyan, Mathur Umang, Bisht Vaishali, Kim Soo Jung, Lee Joshua Junhee, Jin Seung Jung, Siu Yue-Ting, Miele Joshua A, and Yoon Ilmi. 2020. Human-in-the-Loop Machine Learning to Increase Video Accessibility for Visually Impaired and Blind Users. In Proceedings of the 2020 ACM Designing Interactive Systems Conference. 47–60.
- [24]. Yuksel Beste F, Kim Soo Jung, Jin Seung Jung, Lee Joshua Junhee, Fazli Pooyan, Mathur Umang, Bisht Vaishali, Yoon Ilmi, Siu Yue-Ting, and Miele Joshua A. 2020. Increasing Video Accessibility for Visually Impaired Users with Human-in-the-Loop Machine Learning. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 1–9.

Table 1:

Main themes in feedback patterns observed in comments from sighted and blind reviewer. (fill color intensity = frequency)

Themes	Sub-themes	Blind Reviewer (N=560)	Sighted Reviewer (N=560)	Total (N=1120)
Quality	Descriptive	195 (34.82%)	394 (70.36%)	589 (52.59%)
	Objective	64 (11.43%)	14 (2.50%)	78 (6.96%)
	Succinct	8 (1.43%)	1 (0.18%)	9 (0.80%)
	Learning	28 (5%)	48 (8.57%)	76 (6.79%)
	Sufficient	291 (51.96%)	179 (31.96%)	470 (41.96%)
	Interest	58 (10.36%)	0 (0.00%)	58 (5.18%)
	Clarity	267 (47.68%)	24 (4.29%)	291 (25.98%)
	Accurate	85 (15.18%)	20 (3.57%)	105 (9.38%)
	Referable	81 (14.46%)	70 (12.50%)	151 (13.48%)
Speech Act	Instructions	435 (77.68%)	467 (83.39%)	902 (80.54%)
	Question	13 (2.32%)	64 (11.43%)	77 (6.88%)
	Warning	26 (4.64%)	19 (3.39%)	45 (4.02%)
	Compliment	205 (36.61%)	188 (33.57%)	393 (35.09%)
	Required Action	Revision	269 (48.0)	83 (14.82%)
Add information		298 (53.21%)	445 (79.46%)	743 (66.34%)
Fix grammar		56 (10.00%)	0 (0.00%)	56 (5.00%)
Guidance	Suggestion	447 (79.82%)	473 (84.46%)	920 (82.14%)
	Example	308 (55%)	244 (43.57%)	552 (49.29%)
	Clarification	2 (0.36%)	18 (3.21%)	20 (1.79%)