2017
THE JOURNAL ON TECHNOLOGY AND PERSONS WITH DISABILITIES

Eye Movements of Deaf and Hard of Hearing Viewers of Automatic Captions

Kevin Rathbun, Larwan Berke, Christopher Caulfield, Michael Stinson, Matt Huenerfauth

Rochester Institute of Technology

kevinrat@buffalo.edu, larwan.berke@mail.rit.edu, cxc4115@rit.edu, msserd@ntid.rit.edu, matt.huenerfauth@rit.edu

Abstract

To compare methods of displaying speech-recognition confidence of automatic captions, we analyzed eye-tracking and response data from deaf or hard of hearing participants viewing videos.

Keywords

Deaf and Hard of Hearing, Emerging Assistive Technologies, Research and Development

Journal on Technology and Persons with Disabilities
Santiago, J. (Eds): CSUN Assistive Technology Conference

Introduction

Automatic Speech Recognition (ASR) may someday be a viable way to transcribe speech into text to facilitate communication between people who are hearing and people who are deaf or hard of hearing (DHH); however, the output of modern systems frequently contains errors. ASR can output its confidence in identifying each word: if this confidence were visually displayed, then readers might be able to identify which words to trust. We conducted a study in which DHH participants watched videos simulating a one-on-one meeting between an onscreen speaker and the participant. We recorded eye-tracking data from participants while they viewed videos with different versions of this "marked up" captioning (indicating ASR confidence in each word, through various visual means such as italics, font color changes, etc.). After each video, the participant answered comprehension questions as well as subjective preference questions. The recorded data was analyzed by examining where participants' gaze was focused. Participants who are hard of hearing focused their visual attention on the face of the human more so than did participants who are deaf. Further, we noted differences in the degree to which some methods of displaying word confidence led to users to focusing on the face of the human in the video.

Discussion

Researchers have investigated whether including visual indications of ASR confidence helped participants identify errors in a text (Vertanen and Kristensson); later research examined ASR-generated captions for DHH users. In a French study comparing methods for indicating word confidence (Piquard-Kipffer *et al.*), DHH users had a subjective preference for captions that indicated which words were confidently identified. In a recent study (Shiver and Wolfe), ASR generated captions with white text on a black background; less confident words were gray. Several DHH participants indicated that they liked this approach; however, the authors were not

able to quantify any benefit from this confidence markup through comprehension-question testing of participants after they watched the videos. Our study considers captioning to support live-meetings between hearing and DHH participants; so, we investigate ASR-generated captions for videos that simulate such meetings. We display captions in four conditions: no special visual markup indicating ASR word confidence (as a baseline), captions with confident words in yellow color with a bold font, captions with uncertain words displayed in italics, and captions with uncertain words omitted from the text (and replaced with a blank line, e.g. "_____").

A recent study (Sajjad *et al.*) used eye-tracking data to predict how readers would rate the fluency and adequacy of a text. Other researchers used eye-tracking to investigate the behavior of DHH participants viewing videos with captioning, as surveyed in (Kruger *et al.*). Some (Szarkowska *et al.*) found that deaf participants tended to gaze at the caption to read all of the text before moving their gaze back to the center of the video image; whereas hard-of-hearing participants tended to move their gaze back and forth between the captions and the video image, to facilitate speech-reading or use of their residual hearing. Since we are interested in the potential of ASR-generated captions used during live meetings between hearing and DHH participants, it may be desirable to enable the DHH participant to look at the face of their conversational participant as much as possible. For this reason, we analyze the eye-tracking data collected from participants who watched a video that simulates a one-on-one meeting, to examine how much time users are looking at the human's face.

User Study and Collected Data

We produced 12 videos (each approximately 30 seconds) to simulate a one-on-one business meeting between the hearing actor (onscreen) and the DHH viewer. The audio was processed by the CMU Sphinx ASR software (Lamere *et al.*) to produce text output, along with

numerical representation of the system's confidence in each word. This output was used to generate captions for the videos, which appeared at the bottom of the video. The text output had a word-error rate (WER) of approximately 60% depending on the individual video. Figure 1 shows the four display conditions in this study; all participants saw the 12 videos in the sequential order, but the assignment of the four display conditions was randomized for each participant.

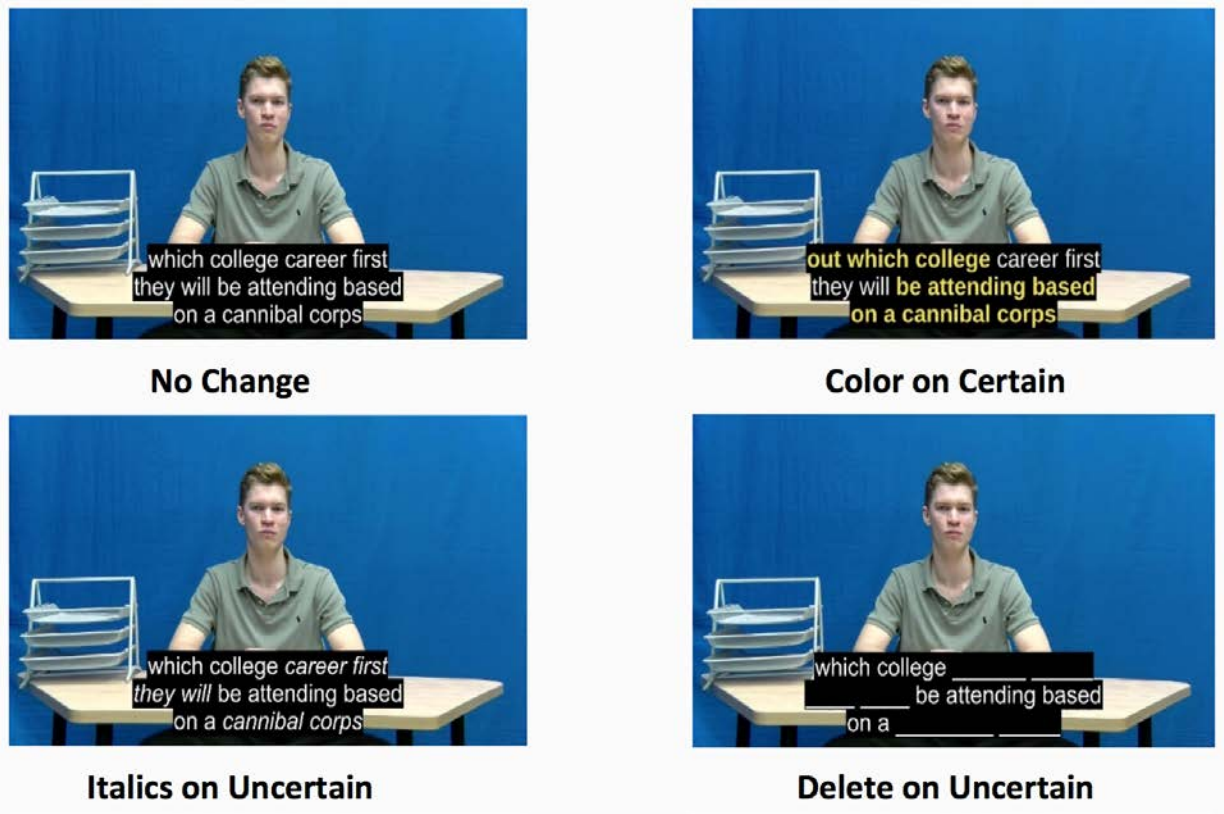

Fig. 1. Image of onscreen stimuli with the four captioning conditions in the study.

Ten participants were recruited using email and social media recruitment on the Rochester Institute of Technology campus: Six participants described themselves as deaf, and four, as hard of hearing. A Tobii EyeX eye-tracker was mounted to the bottom of a standard 23-inch LCD monitor connected to a desktop computer; the eye of the participant was approximately 60cm from the monitor. Software using the Tobii SDK was used to calculate a

list of eye fixations (periods of time when the eyes remain within a defined radius), which include their location on the monitor, along with their start and stop times.

After the arrival of each participant, demographic data was collected, and the eye-tracker was calibrated. After displaying a sample video (to familiarize the participant with the study), all 12 videos were shown (with the sound on, to enable some DHH participants to use residual hearing along with speech-reading, as they might in a real meeting). Eye-tracking data was collected during this initial viewing of the 12 videos. Afterwards, the participants were shown the same 12 videos again, but after viewing each video this second time, participants responded to a Yes/No question asking "Did you like this style of captioning?" Participants also answered multiple-choice questions about factual content conveyed in each video.

Results and Analysis

For eye-tracking data analysis, the onscreen video was divided into several areas of interest (AOI), including (a) the face of the onscreen human and (b) the region of the screen where the captions were displayed, as shown in Figure 2. To analyze the eye-tracking data, we calculate the proportional fixation time (PFT) of that participant on each individual AOI during a video; the PFT is the total time fixated on an AOI divided by the total time of the video. In past studies, time spent fixated on captions usually correlates with the difficulty the reader is having absorbing the content (Robson; Irwin).

Fig. 2. Areas of interest monitored with eye-tracking.

To determine whether the overall patterns of eye-movement recorded in our study were similar to prior work examining the eye-movements of deaf and hard-of-hearing participants, we compared the eye movements of deaf and of hard-of-hearing participants. Significant differences in the "PFT on face" (Mann-Whitney test, $p<0.05$) were found, as shown in Figure 3. This suggests that eye-movement behaviors observed in this study were similar to those observed in prior work with DHH users (Szarkowska *et al*).

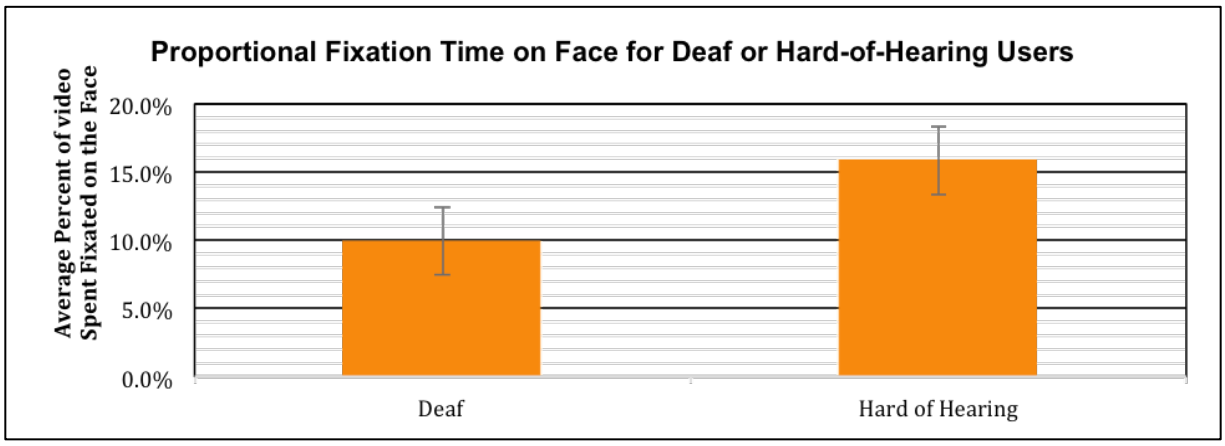

Fig. 3. PFT on Face for Deaf and Hard-of-Hearing Participants.

Figure 4 shows how differences in the display condition also led to differences in participants' time spent looking at the face (Kruskal-Wallis, $p<0.05$); post-hoc Mann-Whitney tests with Bonferroni corrected p-values revealed a significant pairwise difference between the "italics on uncertain" and "delete on uncertain" conditions.

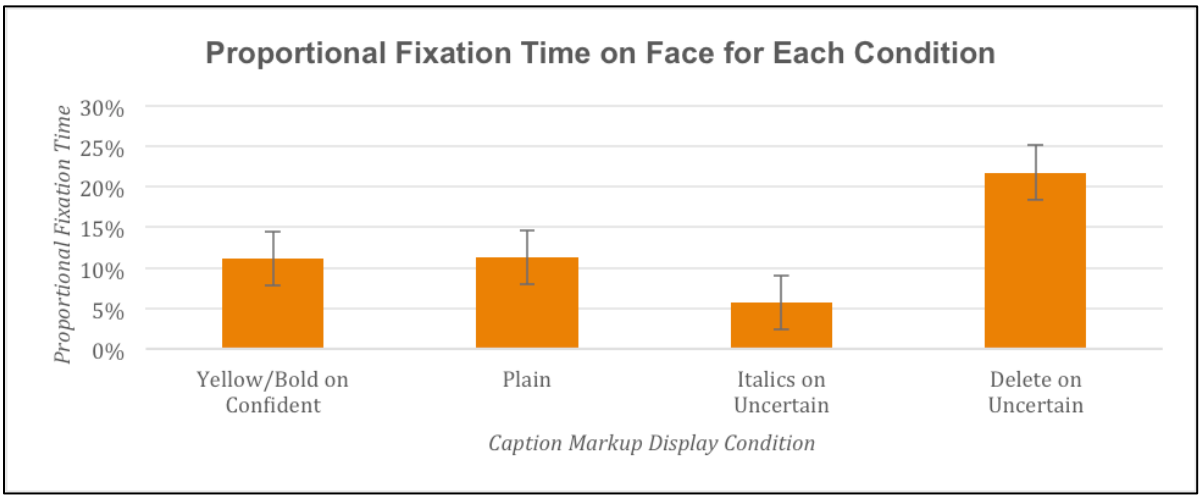

Fig. 4. Percentage of time participants looked at the human's face for each condition.

Participants spent the greatest amount of time looking at the face of the onscreen human when the "delete on uncertain" style of caption display was used. Under the premise that looking at the face of a conversational partner is desirable during a meeting, this might initially suggest that "delete on uncertain" is best. However, we must consider participants' responses to comprehension and preference questions to understand these eye movements. For instance, participants might have spent less time looking at the captions in the "delete on uncertain" condition because they found the captions less useful or simply because there were fewer words to read (since uncertain words were replaced with blank spaces). As indicated in Figure 5, most participants preferred the captions with "italics on uncertain," and as indicated in Figure 6, participants achieved the highest accuracy scores on comprehension questions for captions with

"italics on uncertain." These differences between means, however, were not statistically significant. "Delete on uncertain" had the lowest accuracy scores.

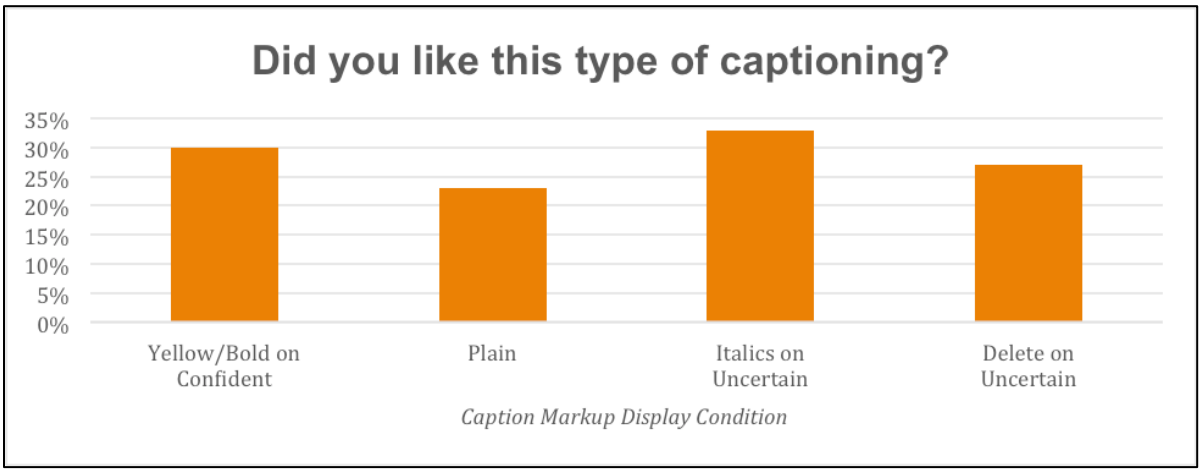

Fig. 5. Subjective preference for each condition.

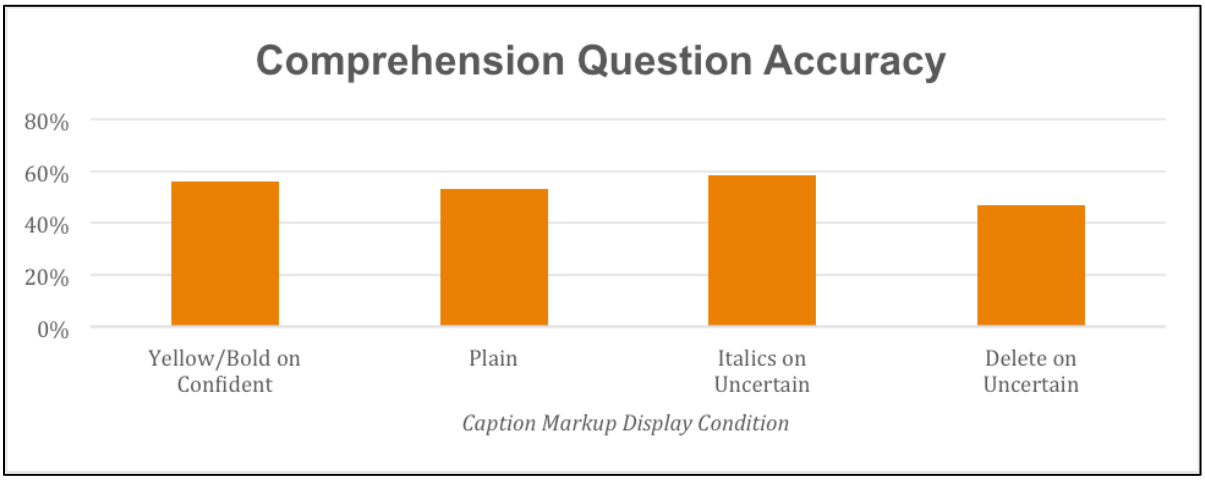

Fig. 6. Comprehension question accuracy for each condition.

Conclusions

This study examined how DHH participants used onscreen captions displayed during videos simulating a one-on-one meeting, with the text of the captions generated using ASR and various conditions of visual presentation of captions to indicate ASR confidence in each word displayed. Eye-tracking analysis revealed that changing the display condition led to differences

in eye-movements of DHH participants. While we initially posited that we should seek to maximize the amount of time that participants look at the face of the human in the video, an analysis of the comprehension and subjective preferences of participants suggests that the relationship between this eye-metric and captioning success is not so straightforward. In future work, we intend to investigate a wider variety of caption display styles and evaluate these approaches with a larger set of participants, to further examine this relationship between eye movements, caption preferences, and methods of displaying confidence in automatic captions.

Works Cited

Irwin, David E. "Fixation location and fixation duration as indices of cognitive processing." In J.M. Henderson & F. Ferreira (Eds.), The interface of language, vision, and action: Eye movements and the visual world, 105-133. New York, NY: Psychology Press. 2004.

Kruger, Jan Louis, Agnieszka Szarkowska, and Izabela Krejtz, "Subtitles on the Moving Image: an Overview of Eye-Tracking Studies" *Refractory* 25. University of Melbourne. 2015.

Lamere, Paul, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. "The CMU SPHINX-4 speech recognition system." *In IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003)*, Hong Kong, vol. 1, pp. 2-5. 2003.

Piquard-Kipffer, Agnès, Odile Mella, Jérémy Miranda, Denis Jouvet, and Luiza Orosanu. "Qualitative investigation of the display of speech recognition results for communication with deaf people." In *6th Workshop on Speech and Language Processing for Assistive Technologies*. 7. 2015.

Robson, Gary D. *The closed captioning handbook*. Amsterdam: Elsevier. 2004.

Sajjad, Hassan, Francisco Guzm´an, Nadir Durrani, Ahmed Abdelali, Houda Bouamor, Irina Temnikova, and Stephan Vogel. "Eyes Don't Lie: Predicting Machine Translation Quality Using Eye Movement." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, June. 2016.

Shiver, Brent, Rosalee Wolfe. "Evaluating Alternatives for Better Deaf Accessibility to Selected Web-Based Multimedia." In Proceedings of the 17th International ACM SIGACCESS

Conference on Computers and Accessibility (ASSETS '15). ACM, New York, NY, USA, 223–230. 2015.

Szarkowska, Agnieszka, Izabela Krejtz, Zuzanna Kłyszejko, Anna Wieczorek. "Verbatim, standard, or edited? Reading patterns of different captioning styles among deaf, hard of hearing, and hearing viewers." *American Annals of the Deaf* 156 (4):363-378. 2011.

Vertanen, Keith, Per Ola Kristensson. "On the benefits of confidence visualization in speech recognition." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1497–1500. 2008.

Journal on Technology and Persons with Disabilities

ISSN 2330-4219

LIBRARY OF CONGRESS * U.S. ISSN CENTER
ISSN Publisher Liaison Section
Library of Congress
101 Independence Avenue SE
Washington, DC 20540-4284
(202) 707-6452 (voice); (202) 707-6333 (fax)
issn@loc.gov (email); www.loc.gov/issn (web page)