Preferred Appearance of Captions Generated by Automatic Speech Recognition for Deaf and Hard-of-Hearing Viewers

Larwan Berke
Rochester Institute of Technology
Rochester, NY, USA
larwan.berke@mail.rit.edu

Khaled Albusays
Rochester Institute of Technology
Rochester, NY, USA
khaled@mail.rit.edu

Matthew Seita
Rochester Institute of Technology
Rochester, NY, USA
mss4296@rit.edu

Matt Huenerfauth
Rochester Institute of Technology
Rochester, NY, USA
matt.huenerfauth@rit.edu

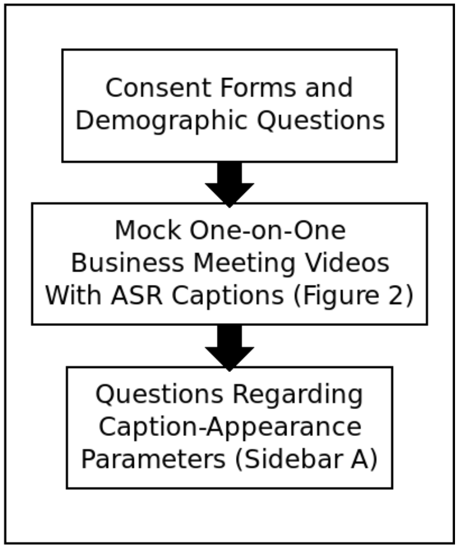

Consent Forms and Demographic Questions
Mock One-on-One Business Meeting Videos With ASR Captions (Figure 2)
Questions Regarding Caption-Appearance Parameters (Sidebar A)

Figure 1: Workflow for our participants during the study.

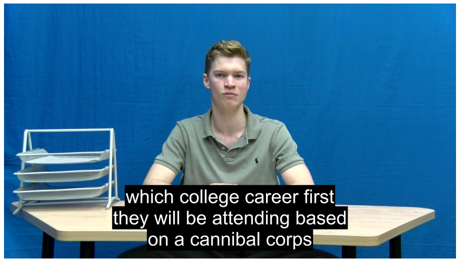


Figure 2: Example video of ASR captions for participants to view during the study.

ABSTRACT

As the accuracy of Automatic Speech Recognition (ASR) nears human-level quality, it might become feasible as an accessibility tool for people who are Deaf and Hard of Hearing (DHH) to transcribe spoken language to text. We conducted a study using in-person laboratory methodologies, to investigate requirements and preferences for new ASR-based captioning services when used in a small group meeting context. The open-ended comments reveal an interesting dynamic between: caption **readability** (visibility of text) and **occlusion** (captions blocking the video contents). Our 105 DHH participants provided valuable feedback on a variety of caption-appearance parameters

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312921

Sidebar A: Caption-appearance Questions for our Participants

(loosely based on prior work [3, 8, 11, 12])

Q1: What type of captioning should we use?
TV style (black box with white captions, see Figure 3), *Movie style* (white text with black outline, see Figure 4), *Movie style* (black text with white outline, see Figure 5), no preference, and other (open-ended textbox).

Q2: How should the captions appear on the screen?
TV CC style (one word at a time), *Movie Subtitle style* (entire line at a time), no preference, and other.

Q3: Where should the captioning be located?
Inside the video (bottom), Inside the video (top), Outside the video (below), Outside the video (above), Outside the video (right), Outside the video (left), no preference, and other.

Q4: How many lines of captioning should be shown on the screen?
1, 2, 3, 4, 5, no preference, and other.

Q5: Which font do you prefer when viewing captions?
Arial, Comic Sans, Copperplate, Courier, Droid Sans Mono, Georgia, Helvetica, Monotype Corsiva, Times New Roman, Tiresias, Verdana, no preference, and other. (see Figure 6)

(strongly preferring **familiar styles** such as closed captions), and in this paper we start a discussion on how ASR captioning could be visually styled to improve text readability for DHH viewers.

CCS CONCEPTS

- **Human-centered computing** → **Empirical studies in HCI**; *Accessibility systems and tools*; **User interface design**.

KEYWORDS

Automatic Speech Recognition; Captioning; Deaf and Hard-of-Hearing; Appearance; User Interface.

INTRODUCTION

People who are Deaf and Hard of Hearing (DHH) make use of a wide variety of communication technologies and accommodations, e.g. real-time captioning services produced by professional transcriptionists (with text displayed on a screen for the user) or American Sign Language (ASL) interpreting [13]. Furthermore, DHH individuals who do not identify as culturally Deaf or older adults who have lost hearing later in life may prefer text-based accessibility tools [7], rather than sign language interpretation.

Recent breakthroughs in Automatic Speech Recognition (ASR) wherein ASR is nearing human-level accuracy could enable communication tools such as ASR transcribing speech to text, for DHH individuals on their mobile devices, with the assistance of cloud-based services. In small-group meetings with colleagues who are hearing, DHH users might view live captions generated by ASR tools, as seen in a prototype visualized in Figure 2.

Since it is known that ASR technologies are currently **imperfect** [1, 4, 10], DHH users may have different preferences for the display of captions from ASR than when compared to captions from human transcribers. Prior work has investigated a variety of caption/subtitle appearance options, but very few asked participants questions about how the captions should be styled. However, **none** of the prior work looked at the intersection of ASR and caption-appearance preferences for small-group meetings (in-person or remotely via technologies such as Skype) with DHH individuals.

Prior Work On ASR Captioning

Researchers [15] have considered whether ASR could provide text transcriptions of spoken language for DHH users. Elliot *et al.* [4] revealed in an exploratory survey of DHH users that they were agreeable to having ASR support their conversations at the workplace. Likewise, Kawas *et al.* [10] and Berke *et al.* [1] discovered that the DHH community was receptive to ASR captioning of spoken information in classrooms and small group meetings.

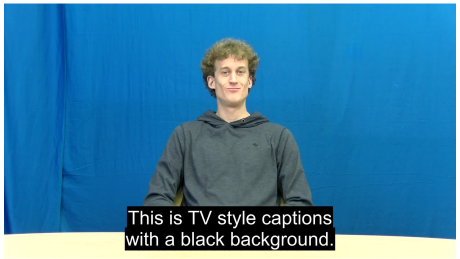


Figure 3: (Q1-Type) Typical "TV-style" captions with white text and a black box.

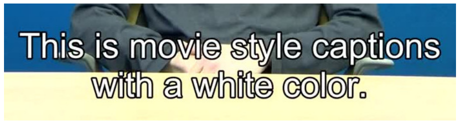


Figure 4: (Q1-Type) "Movie-style" captions with white text and black outline.

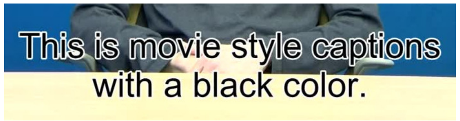
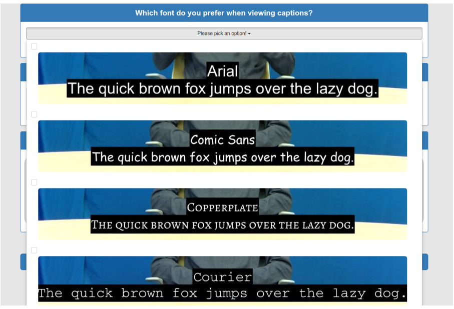


Figure 5: (Q1-Type) "Movie-style" captions with black text and white outline.



Which font do you prefer when viewing captions?

Please pick an option!

□ Arial: The quick brown fox jumps over the lazy dog.

□ Comic Sans: The quick brown fox jumps over the lazy dog.

□ COPPERPLATE: THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG.

□ Courier: The quick brown fox jumps over the lazy dog.

Figure 6: (Q5-Font) Example of choices.

Other researchers investigated how to display the captions for viewers in a variety of styles and methods. Tracked captions (captioning projected above the speaker and following their movement) were found by Kushalnagar *et al.* [12] to enhance the perception of captions by DHH users in classrooms. Crabb *et al.* [3] discussed the UX of subtitle position for online videos and made several recommendations on how to style the captions (manually created by human transcriptionists). Gower *et al.* [5] looked into using speech pauses for automatic punctuation of ASR captions. Eye-tracking was used by Szarkowska *et al.* [16] to monitor how DHH users read different styles of captioning (verbatim, standard, or edited captions). Hong *et al.* looked at how captioning could be moved dynamically inside the video frame to improve the accessibility for DHH users [6]. Finally, Peng *et al.* [14] investigated placement of speech caption bubbles on Augmented Reality devices.

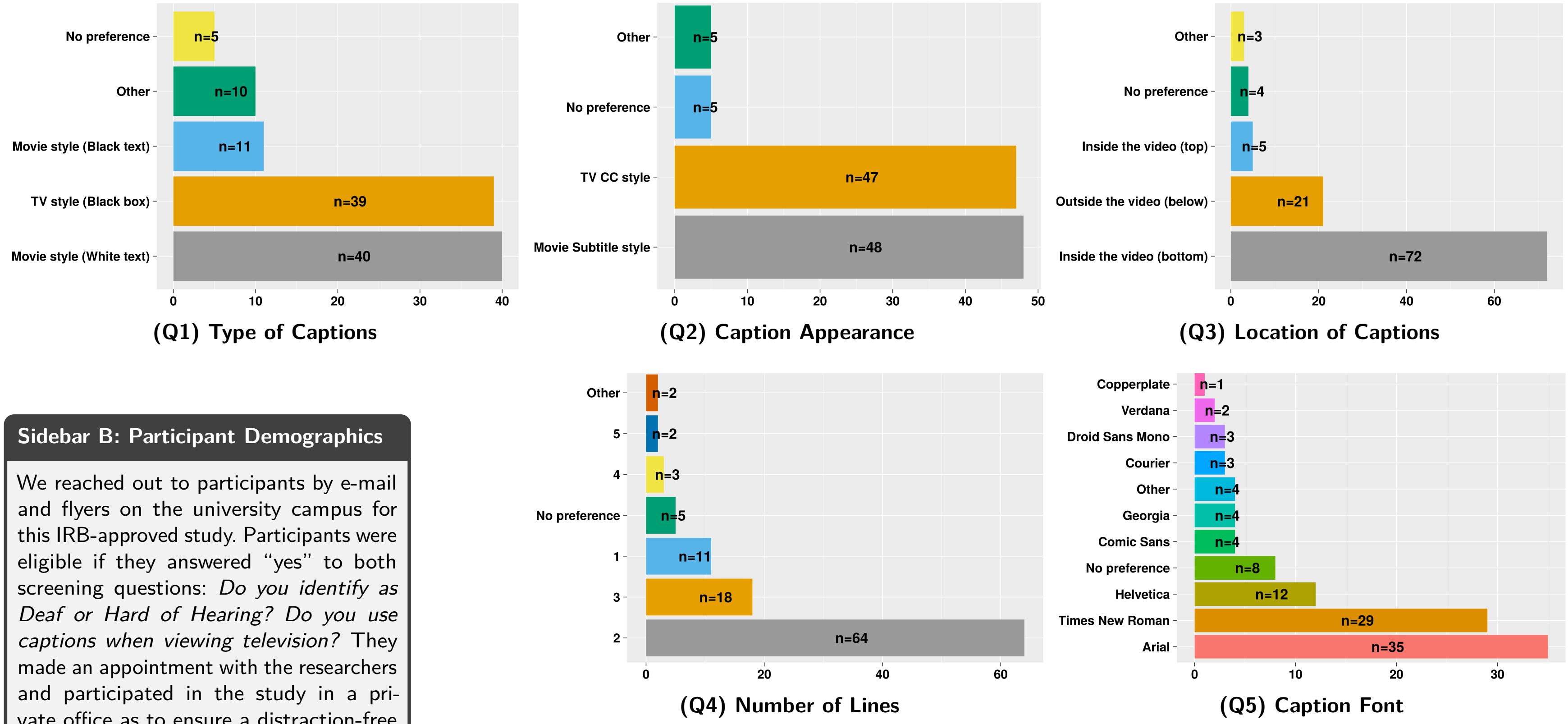
METHODOLOGY

We conducted an in-person controlled experiment at our laboratory with DHH participants (see Sidebar B for information on the participant demographics), as **part of a larger project** which studied how DHH users might use ASR tools [1, 2, 9]. After completing consent forms and a demographic questionnaire, our participants were shown a video to introduce the business meeting scenario which simulated the experience of a participant engaging in a meeting with a hearing individual, with the aid of automatic captions (see Figure 1 for an outline of the participant workflow). Participants were informed that the words they would see in the captions were produced by a computer that was trying to identify what was spoken automatically, and that errors would appear from time to time as ASR is still imperfect.

Participants then viewed sample videos of ASR captioning a mock business meeting (with an average Word Error Rate of 23.2%), such as seen in Figure 2 wherein a speaker is sitting behind a desk with an example of ASR captions: "*which college career first they will be attending based on a cannibal corps*". The speaker actually said: "which college career fairs they will be attending based on the candidate requirements." Participants were given several videos to get acquainted with the idea of ASR captioning before moving on to the next phase of the experiment wherein we gathered their opinion on the user-interface parameters.

Q1-5 in Sidebar A contains a list of questions our research team gave to our DHH participants. For each question, we created an ASL video describing the question and answer options with closed captions. Several questions such as the font choice also had pictures of the options as to reduce participants' cognitive burden when selecting the answers (see Figure 6). Finally, we gave participants the opportunity to express any feedback and opinions they had about captioning parameters via open-ended questions such as "*Do you have any comments on how the captions should appear?*".

RESULTS



(Q1) Type of Captions

(Q2) Caption Appearance

(Q3) Location of Captions

(Q4) Number of Lines

(Q5) Caption Font

Figure 7: Results from the study for our 5 questions (N=105).

A chi-square test of goodness-of-fit was performed to determine whether some answer choices were preferred over others, and all questions had **significant** results: [**Q1** χ^2=55.333, df=4, p=2.766e^{-11}], [**Q2** χ^2=68.829, df=3, p=7.604e^{-15}], [**Q3** χ^2=165.24, df=4, p<2.2e^{-16}], [**Q4** χ^2=200.53, df=6, p<2.2e^{-16}], and [**Q5** χ^2=240.67, df=10, p<2.2e^{-16}]. Answer choices significantly preferred over the rest include: **Q1 (Type):** *Movie style (white text)* and *TV style (black box)*, **Q2 (Appearance):** *Movie subtitle style* and *TV CC style*, **Q3 (Location):** *Inside the video (bottom)*, **Q4 (Number of Lines):** *2 lines*, and **Q5 (Font):** *Arial* and *Times New Roman*.

Sidebar B: Participant Demographics

We reached out to participants by e-mail and flyers on the university campus for this IRB-approved study. Participants were eligible if they answered "yes" to both screening questions: *Do you identify as Deaf or Hard of Hearing? Do you use captions when viewing television?* They made an appointment with the researchers and participated in the study in a private office as to ensure a distraction-free environment. Participants were paid $40 for the 60-minute study. A total of **105 DHH individuals** participated, and they self-identified their hearing status as (69 Deaf, 36 Hard-of-Hearing), and gender as (58 males and 47 females). Participants' ages ranged from 18-30 years old [mean=22.105, median=22].

Sidebar C: Quotes from Participants

The authors (two are DHH) conducted a thematic analysis of 116 comments (2069 words) using a two-round consensus coding process. Some of the top themes were:

Familiarity With Specific Style (n=43)

– *common like movie caption* (P18-Q4)

– *... read the same location from anywhere else such as TV Netflix Movie Youtube video game. keep that way* (P20-Q3)

– *I like TV CC to read flow in order of the word I don't like the movie box at once pop up with multiple words that I might possibly miss the words.* (P7-Q2)

Readability of Captions (n=24)

– *It's better to leave the black box because it helps me to read better* (P5-Q1)

– *The movie style with white text are the best. It helps by standing out what is behind the texts ...* (P21-Q1)

Captions Occluding Video (n=21)

– *Love movie style because you can see more picture!* (P92-Q1)

– *Should be 2 lines otherwise it will take up the space in the video and the audience wouldn't know if the video plays an important role as well or not.* (P68-Q4)

Captions Closer to Speaker (n=6)

– *... the captioning should be above the speaker (or top) whenever possible so the words are closest to person's face to help us understand what the person is expressing tone/emotional ...* (P21-Q3)

– *The captioning at the top would be good idea since you can see the actions or people speaking to you so you can read the captioning.* (P73-Q3)

DISCUSSION

Both **Q1 (Captioning Type) and Q2 (Caption Appearance)** revealed that our participants were almost divided in their preference for traditional "TV CC Style" or the "Movie Subtitle Style" captions. For **Q3 (Location of Captions)**, our participants strongly preferred to see captions inside the video (bottom), which contrasted with prior work from Crabb *et al.* [3] which recommended the captions be outside of the video frame for online videos. In **Q4 (Number of Lines)** our participants strongly preferred to see 2 lines of captions, whereas Kushalnagar *et al.* [12] observed a preference for 3 or more lines. We speculate that these differences may be due to our study's focus on captions for small-group live meetings, rather than for viewing online videos (as in prior work).

Our thematic analysis of the open-ended comments (see Sidebar C for some quotes from our participants) revealed a **tension** between caption **readability** and captions **occluding** the video frame (for example, the black box helps reading the text but hinders visibility of the video contents). Participants expressed a desire for familiarity in caption appearance (TV CC or Movie Subtitle) style, yet they also expressed a desire for the ability to customize the captioning, to suit the particular ambience and/or background of the video, to improve readability or to reduce occlusion.

CONCLUSION AND FUTURE WORK

All of our questions (**Q1-5**) had participants whom preferred each potential option, which showed the diverse perspectives in the DHH community. Some users had a strong preference for "traditional" captioning as seen on televisions, yet others preferred the movie subtitle style, and others wanted entirely novel combinations of options. One participant summarized well the potential contribution of ASR to **small-group meetings**: *"it will help deaf students while they are doing interview process, work, restaurants, court, during eat meals with friends and families."* (P27-Q3) We strongly recommend programmers incorporating ASR captioning into their software **empower their users** with the ability to customize the styling and appearance of the captions as much as possible, in order to reduce the viewer's cognitive burden when reading the captions as one participant said it well: "*I think giving the preference options is the best way you can do*" (P66-Q2).

Our team is currently conducting a second study exploring this problem space (with a larger pool of online participants) in order to investigate our speculation as to whether there are differences in preference when DHH users utilize ASR for small-group meetings or online videos (e.g. YouTube).

ACKNOWLEDGMENTS

We thank Christopher Caulfield, Kasmira Patel, Anmolvir Kaur, and our study participants. This material is based upon work supported by the National Science Foundation under Award No. 1746056, by the Department of Health and Human Services under Award No. 90DPCP0002-01-00, by a Google Faculty Research Award, and by the National Technical Institute of the Deaf (NTID).

REFERENCES

[1] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. 2017. Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One Meetings. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery (ACM), New York, NY, USA, 155–164. https://doi.org/10.1145/3132525.3132541

[2] 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. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '18)*. Association for Computing Machinery (ACM), New York, NY, USA, Article 91, 12 pages. https://doi.org/10.1145/3173574.3173665

[3] 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 and Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery (ACM), New York, NY, USA, 215–222. https://doi.org/10.1145/2700648.2809866

[4] Lisa Elliot, Michael Stinson, James Mallory, Donna Easton, and Matt Huenerfauth. 2016. Deaf and Hard of Hearing Individuals' Perceptions of Communication with Hearing Colleagues in Small Groups. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery (ACM), New York, NY, USA, 271–272. https://doi.org/10.1145/2982142.2982198

[5] 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 (ASSETS '18)*. ACM, New York, NY, USA, 414–416. https://doi.org/10.1145/3234695.3241023

[6] Richang Hong, Meng Wang, Mengdi Xu, Shuicheng Yan, and Tat-Seng Chua. 2010. Dynamic Captioning: Video Accessibility Enhancement for Hearing Impairment. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery (ACM), New York, NY, USA, 421–430. https://doi.org/10.1145/1873951.1874013

[7] Marylyn Howe and Bill Graham. 1990. The importance of captioning for late-deafened adults. *International Journal of Technology & Aging* 3, 2 (1990), 121–131.

[8] Carl Jensema, Ralph McCann, and Scott Ramsey. 1996. Closed-Captioned Television Presentation Speed and Vocabulary. *American Annals of the Deaf* 141, 4 (1996), 284–292. http://www.jstor.org/stable/44401017

[9] Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the Usability of Automatically Generated Captions for People Who Are Deaf or Hard of Hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery (ACM), New York, NY, USA, 165–174. https://doi.org/10.1145/3132525.3132542

[10] Saba Kawas, George Karalis, Tzu Wen, and Richard E. Ladner. 2016. Improving Real-Time Captioning Experiences for Deaf and Hard of Hearing Students. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. Association for Computing Machinery (ACM), New York, NY, USA, 15–23. https://doi.org/10.1145/2982142.2982164

[11] Jan-Louis Kruger, Agnieszka Szarkowska, and Izabela Krejtz. 2015. Subtitles on the moving image: An overview of eye tracking studies. *Refractory* 25 (2015), 14. http://refractory.unimelb.edu.au/2015/02/07/kruger-szarkowska-krejtz/

[12] Raja S. Kushalnagar, Gary W. Behm, Aaron W. Kelstone, and Shareef Ali. 2015. Tracked Speech-To-Text Display: Enhancing Accessibility and Readability of Real-Time Speech-To-Text. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers and Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery (ACM), New York, NY, USA, 223–230. https://doi.org/10.1145/2700648.2809843

[13] Michella Maiorana-Basas and Claudia M. Pagliaro. 2014. Technology Use Among Adults Who Are Deaf and Hard of Hearing: A National Survey. *The Journal of Deaf Studies and Deaf Education (JDSDE)* 19, 3 (1 July 2014), 400–410. https://doi.org/10.1093/deafed/enu005

[14] Yi-Hao Peng, Ming-Wei Hsi, Paul Taele, Ting-Yu Lin, Po-En Lai, Leon Hsu, Tzu-chuan Chen, Te-Yen Wu, Yu-An Chen, Hsien-Hui Tang, and Mike Y. Chen. 2018. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '18)*. Association for Computing Machinery (ACM), New York, NY, USA, Article 293, 10 pages. https://doi.org/10.1145/3173574.3173867

[15] Soraia Silva Prietch, Napoliana Silva de Souza, and Lucia Villela Leite Filgueiras. 2014. A Speech-To-Text System's Acceptance Evaluation: Would Deaf Individuals Adopt This Technology in Their Lives?. In *Proceedings of the 8th International Conference on Universal Access in Human-Computer Interaction* (Heraklion, Crete, Greece) *(UAHCI '14: Design and Development Methods for Universal Access)*, Constantine Stephanidis and Margherita Antona (Eds.). Springer International Publishing, Cham, 440–449. https://doi.org/10.1007/978-3-319-07437-5_42

[16] Agnieszka Szarkowska, Izabela Krejtz, Zuzanna Klyszejko, and Anna Wieczorek. 2011. 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 (2011), 363–378. https://doi.org/10.1353/aad.2011.0039