Towards Improving Real-Time Head-Worn Display Caption Mediated Conversations with Speaker Feedback for Hearing Conversation Partners

Jenna Jiayi Kang
jkang394@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Emily Layton
elayton7@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

David Martin
dmartin305@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Thad Starner
thad@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Figure 1: Left: Hearing participant cannot see captions. Center: Hearing participant can see captions. Right: tooz DevKit head-worn display.

ABSTRACT

Many products attempt to provide captioning for Deaf and Hard-of-Hearing individuals through smart glasses using automatic speech recognition. Yet there still remain challenges due to system delays and dropouts, heavy accents, and general mistranscriptions. Due to the imperfections of automatic speech recognition, there remains conversational difficulties for Deaf and Hard-of-Hearing individuals when conversing with hearing individuals. For instance, hearing conversation partners may often not realize that their Deaf or Hard-of-Hearing conversation partner is missing parts of the conversation. This study examines whether providing visual feedback of captioned conversation to hearing conversation partners can enhance conversational accuracy and dynamics. Through a task-based experiment involving 20 hearing participants we measure the impact on visual feedback of captioning on error rates, self-corrections, and subjective workloads. Our findings indicate that when given visual feedback, the average number of errors made by participants was 1.15 less ($p = 0.00258$) indicating a notable reduction in errors. When visual feedback is provided, the average number of self-corrections increased by 3.15 ($p < 0.001$), suggesting a smoother and more streamlined conversation These results show that the inclusion of visual feedback in conversation with a Deaf or Hard-of-Hearing individual can lead to improved conversational efficiency.

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650976

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in accessibility**; **Empirical studies in ubiquitous and mobile computing**.

KEYWORDS

Deaf, Hard-of-Hearing, head-worn display, captioning, accessibility

ACM Reference Format:
Jenna Jiayi Kang, Emily Layton, David Martin, and Thad Starner. 2024. Towards Improving Real-Time Head-Worn Display Caption Mediated Conversations with Speaker Feedback for Hearing Conversation Partners. In *Extended Abstracts of the CHI Conference on Human Factors in Computing*

Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3613905.3650976

1 INTRODUCTION

The World Health Organization estimates 25% of the world will have some degree of hearing loss by 2050 [5]. Current medical interventions may include the use of hearing aids. However, this fails to restore full hearing as well as speaker localization [7]. Furthermore, hearing aids are unaffordable for 75% of people with hearing loss in the United States as they typically cost between $1000-$6000 [14]. Another method many use is Automatic Speech Recognition **(ASR)** on mobile devices or laptops [11] with applications such as Google's Live Transcribe [10]. However, while ASR has made great leaps in the last few years, it still is prone to mistranscriptions which may slow and increase miscommunications in ASR-mediated conversations.

Leveraging the progress in ASR, Head-Worn Display **(HWD)** captioning systems has been growing as a research area for the last several years [3, 8, 12, 15, 19] and continues to grow as HWDs form factors become lighter. Such investigations include HWD captioning systems in mobile contexts [12], field-of-view [3], adjustable fonts [8], indicators for sound localization [13], user interfaces for displaying captions and other contextual information [3, 13], and social acceptability of wearable usage by individuals with disabilities [20]. However, HWD-based captioning systems that are capable of everyday usage with a discrete and comfortable form factor have yet to hit the market at scale as there are many design factors in need of investigation. There exists a significant amount of literature on providing communication access to Deaf and Hard-of-Hearing **(DHH)** individuals via HWD-based captioning systems but little on how to provide hearing conversational partners with ways that they can improve the conversation.

Through our investigations, the following research questions are pursued:

- **Does having feedback of captions for the hearing partner in a conversation with a DHH person improve the captioning experience?**
- **Does having feedback on captions reduce the number of conversational errors?**

1.1 Motivation and Related Work

Traditionally, real-time captioning services are provided by a human stenographer or transcriptionist. However, such services are not suitable for spontaneous use since they require scheduling in advance. Furthermore, it can be costly at over 150 dollars an hour. ASRs advancements in the last few years, low cost, and robustness has prompted research into its integration into the day-to-day life of DHH individuals. Recently, we have seen ASR integrated into HWDs [3, 12, 13, 17, 19, 22], the classroom [4, 23] and videoconferencing [2, 16]. Additionally, Human-Assisted ASR has been explored via crowdsourcing in university lecture slides [23] and educational videos [6]. When developing accessible technologies for DHH and disabled individuals it is important to consider how socially acceptable its use is for bystanders [9, 18, 20, 21], and a recent publication by Olwal et al. [17] emphasize this need for HWD captioning as well, going so far as to suggest the captioning glasses be virtually indistinguishable from normal eyeglasses.

However, once such devices are possible, the technology hidden behind the curtains can become unknown, and the hearing conversation partner might take for granted the accuracy of the ASR systems. We study the interactions, both quantitatively and qualitatively, in HWD-based ASR mediated conversations between DHH and hearing conversation partners. Furthermore, we hypothesize that providing visual caption feedback will lead to more efficient conversations, lower mistranscriptions, and thus removing the burden of requesting hearing speakers to repeat themselves.

2 COMPARING CONVERSATIONAL EXPERIENCE WITH AND WITHOUT CAPTIONING FEEDBACK TO THE HEARING PARTICIPANT

Following previous work, this study uses the tooz Dev-Kit HWD [22] and ToozKit, an open-source application for real-time captioning on Android-based HWDs [8]. The ToozKit application utilizes the microphone of a mobile device, which also transcribes and transmits captions via Bluetooth to the tooz Dev-Kit HWD[1]. A Samsung Galaxy Z Flip5 is used to display transcriptions to the hearing participant and the tooz smart glasses are used for displaying transcriptions to the researcher acting as a DHH individual. An iPad is used to write the information relayed to the researcher by the participant during the study (see Figure 1).

The task emulates visiting the Department of Motor Vehicles (DMV) to renew a driver's license where the DMV employee is deaf[2]. This task was chosen to ensure that participants could complete a task with information that is easy to articulate to the researcher. A form for license renewal was created based on the researcher's state DMV license renewal form (Figure 3). Six task cards were created with random names, numbers, addresses, and attributes that were randomly chosen throughout all of the trials without replacement (Figure 2). All trials were conducted in the same quiet room, with the participant and DMV worker (researcher) sitting on two sides of a table, facing each other, and a second researcher sitting to the side to track metrics in real time.

There is an important distinction here: given that our research primarily focuses on experimenting with hearing participants rather than DHH users, having the researcher act as deaf and rely solely on transcriptions on the HWD system aligns with our experimental goals. The fact that they are not deaf does not affect our study, since our target population consists of hearing individuals. Furthermore, we note that a DHH researcher assisted, consulted, and contributed every step of the way in this study from ideation to the write-up of this paper.

2.1 Participants

All participants were screened through a quick survey on their experience with HWDs and ASR systems. No participants had experience on HWD-based captioning systems, two had experience

[1] https://tooz.com/devkit/

[2] We use 'deaf' to refer to the audiological status, whereas 'Deaf' refers to being a member of the Deaf Community

Information Card #1

Disclaimer: We will not use any personal information as a part of this experiment. Any personal information given (such as address and birthday) will be immediately discarded afterwards.

Email: yanfangate@gmail.com
Phone: 202-334-5570

No georgia license number

First Name: Yanfang
Middle Name: Lake
Last Name: Escalus
Maiden Name: None
Social Security: 334-98-2003
Suffix: Jr.

Address: 2 South Talbot St.
Zip: 32714
City: Altamonte Springs
State: North Carolina

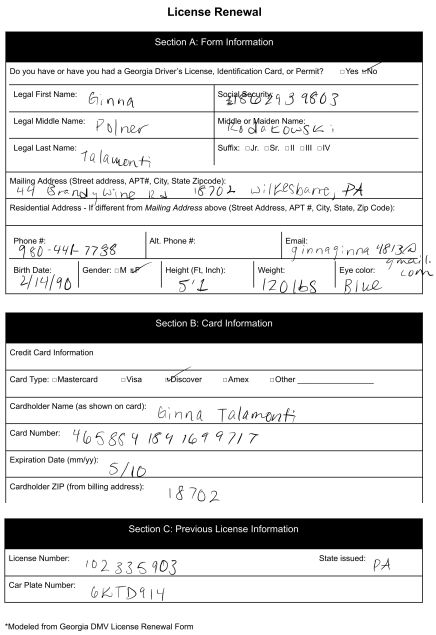
Birth Date: March. 8, 1997
Weight: 155 lbs
Eye Color: Blue

Gender: Male
Height: 6'1

Credit Card Type: Amex
Card Holder Name: Yanfang Escalus
Credit Card Number: 3735-9900-5095-005
Expiration Date: 6/26
Cardholder ZIP: 27514

License Number: 626-926-253
State Issued: Montana
Car Plate Number: 668450B

Figure 2: One of the task cards provided to participants depicting basic demographic information and payment information needed for a license renewal.



*Modeled from Georgia DMV License Renewal Form

Figure 3: An example of the license renewal forms filled out by the research acting as a DMV clerk.

interacting with DHH individuals, and few had experience using HWDs.

20 participants were recruited through snowball sampling who were on average 21.4 years old ($\sigma = 2.62$); 10 self-reported as male and 10 as female; spoke an average 1.8 languages ($\sigma = 0.616$) across nine distinct languages as seen in Table 1 in Appendix A.

2.2 Methods

The study design is AB/BA format where participants were unable to see the transcriptions as they went through one of the tasks (A, "control") and were able to see the transcriptions the worker saw on the device in the other task (B, "feedback"). These scenarios were alternated so odd-numbered participants were AB and even-numbered participants were BA. In task A, the Samsung Galaxy Z Flip5 was placed to the left of the participant, close enough to ensure that their voice would be picked up for transcriptions. In task B, participants were told to place the device wherever they felt most comfortable to view the transcriptions. Our independent variable is the availability of transcriptions.

2.3 Metrics

In conjunction with the data collected through the NASA Task Load Index (TLX) and System Usability Scale (SUS) assessments, additional metrics were included to further analyze our experiments. These metrics provide an understanding of the communication dynamics during the experiments, shedding light on the number of errors, the necessity for repeated information to the DHH individual, and participant-driven corrections, in control vs. feedback experiments.

2.3.1 Incorrect Characters. The number of incorrectly written characters was recorded relative to the information provided on the task card. This metric serves as an indicator of lower-level accuracy in the execution of the task.

2.3.2 Prompted Corrections. The second (and subsequent) occurrences where the DMV worker (researcher) had to request the same information from the participant are called "prompted corrections." These track the need for repetition or clarification in the communication process.

2.3.3 Unprompted Corrections. When participants voluntarily repeat themselves without the DMV worker prompting for further information. This metric differentiates between one-sided fixes initiated by the participant and captures instances where participants independently recognized and rectified the transcription errors.

3 STUDY

3.1 Procedure

Each session took around forty-five minutes to complete. To begin, participants were given brief instructions and a task card with information to examine. Participants were told that they would be visiting the DMV to renew a license and that the worker asking them questions (the researcher) was a deaf individual who could only understand them by reading captions from their HWD. Participants were tasked with answering all of the worker's questions based on the information given on the randomly selected task card. The researcher was also previously acquainted with the technology in unofficial runs of this procedure, to fully, and solely, rely upon the captions for the purpose of the study.

Once the task began (either control or feedback), the worker would ask the participant questions to complete the license renewal form. The worker only wrote information they could see on the HWD while another researcher collected prompted and unprompted corrections throughout each session. After the first session ended, the participant completed a NASA TLX and SUS (on a 7-point scale) survey to gauge how they perceived their workload and the system presented. The next session was then conducted with the reversed AB format followed by another set of NASA TLX and SUS surveys. Concluding the entirety of the study, the participant was verbally asked a few semi-structured interview questions (in Appendix B) to better understand how they felt about both sessions and if they had a preference for either.

There was further data analysis after the completion of each trial, where the number of mistakes was recorded relative to the original task information card.

4 RESULTS

4.1 System Usability Survey (SUS)

Using a paired t-test, the difference in mean SUS scores did not reach statistical significance ($p = 0.285$). Average control condition score was 81.81; average feedback condition score was 76.85. Both are considered to be "good," but not "excellent" scores [1].

4.2 NASA TLX

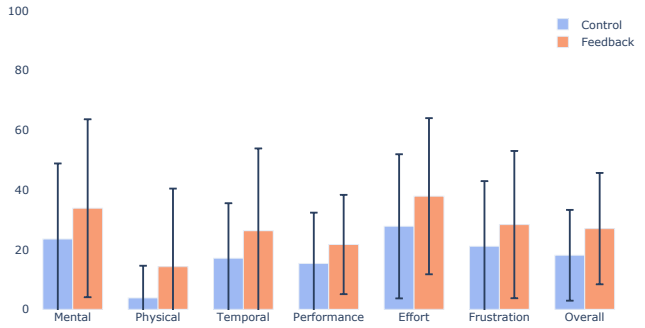


Figure 4: Average NASA TLX scores with and without caption feedback on the phone. The six set of bars represent the different subscales with the total average of the six subscales on the far right.

Overall, there was a significant difference between the overall workload average between the control and feedback trials **(p=0.020)**. The average difference in workload between the feedback and control conditions was 8.958 (with feedback having a higher workload).

During control trials, participants reported an overall workload (averaging the six scales) of 17.92 with a standard deviation of 17.21. Participants' average ratings were mental (μ =23.21, σ =29.52), physical (μ =1.07, σ =2.89), temporal (μ =18.93, σ =22.03), performance (μ =17.14, σ =20.54), effort (μ =30.00, σ =26.53), and frustration (μ =17.14, σ =22.08).

In feedback trials, participants reported an overall average of 28.57 with a standard deviation of 21.16. Participants' average ratings were mental (μ =32.50, σ =32.21), physical (μ =17.50, σ =31.73), temporal (μ =31.43, σ =32.61), performance (μ =21.07, σ =16.43), effort (μ =41.43, σ =26.85), and frustration (μ =27.50, σ =26.66). See Figure 4.

4.3 Mistakes, Corrections (prompted and unprompted), and Task Time

In general, the average number of mistakes made was significantly lower in feedback conditions **(p=0.0258)**. An order effect was apparent. Participants who completed the feedback trial first (BA order), made fewer mistakes in the control condition due to their adaption to the captions during the feedback trial.

During control trials, the average number of mistakes made was 3.6 (σ = 0.33). This decreased by 1.15 **(p=0.026)** in feedback trials (μ = 1.5, σ = 3.39). The average number of prompted corrections showed statistically significant higher results for control trials, with a difference of 1.85 **(p=0.004)**. Unprompted corrections also showed statistical significance, indicating higher unprompted corrections in feedback trials, with a difference of 3.15 **($p < 0.001$)**. See Figure 5.

All participants were able to complete the entire task, with an average time of 413 seconds. Via a paired t-test, the average time to complete the feedback trial (μ = 428.9, σ = 125.8) was significantly higher than the average time to complete the control trial (μ = 398.5, σ = 91.1) **(p=0.0138)**. The average time taken to complete the task increased by 30.9 seconds during feedback trials, in comparison to control trials.

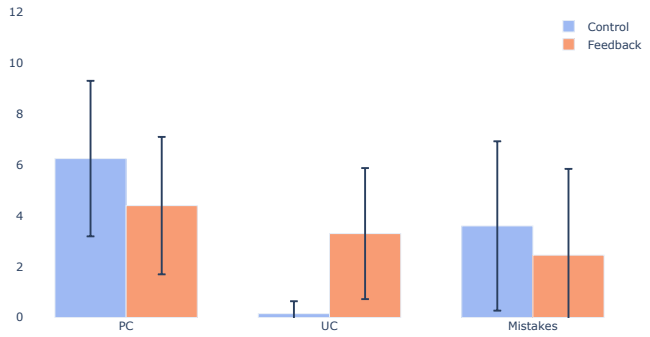


Figure 5: Average Prompted corrections (PC), unprompted corrections (UC), and mistakes for with and without caption feedback on the phone.

5 DISCUSSION

Aligning with our intuition, there were significantly fewer mistakes made in feedback trials than in control trials. Additionally, there were more self-corrections in feedback trials than in control trials, signifying less back-and-forth between the participant and researcher for a given entry in the task. In contrast, there were significantly more prompted corrections (initiated by the researcher) in control trials, due to the lack of self-corrections.

Comparing the time required to complete the task between feedback and control conditions reveals that the feedback condition had significantly longer durations. This result suggests a potential

connection to a more mentally strenuous task during the extended feedback trials compared to the shorter control trials. Analyzing participant interviews, seven participants said they would prefer to use the system without feedback as it required less cognitive workload. One participant (P4) reported feeling more annoyed when they could see feedback because they had to focus more on the transcriptions and less on their conversation with the worker.

While the overall results support intuition, this finding was an interesting outcome of the study as, for some participants, it may indicate a lack of empathy from the hearing individual on the cognitive load placed on the DHH in spoken conversation.

Based on the data and interviews conducted, three key themes arose.

5.1 Theme 1: Impact to Natural Conversational Behaviors

Although it seemed like participants' responses to using the system came from a potential lack of empathy towards individuals who are Deaf or Hard of Hearing (DHH), participants reported not liking the feedback system because of the effects of impersonal conversational behaviors. They felt that using the feedback system drew their attention away from the worker and perceived that as a negative effect on the conversation. Further investigation revealed that some participants showed a desire for more face-to-face communication through eye contact and direct verbal interaction.

P1: Seeing the transcriptions would be better for me, but wouldn't be better for the deaf or hard of hearing person. I would be too focused on correcting the transcriptions and not on actually speaking with the person.

P3: Seeing the captions was very distracting and I lacked eye contact.

P14: I think it was easier without the transcriptions because I could focus on the conversation and getting correct cues from the DHH instead of looking at the captioning the whole time.

P19: I preferred not seeing them [transcriptions] because the conversation was more natural instead of me being conscious of the system.

As many participants relied on confirmation and feedback from the worker, not being able to give them their full attention was a big factor for them when debating the usefulness of the feedback system.

5.2 Theme 2: Changes to Speaking Behaviors

Many participants changed their behaviors as they spoke to the worker in both control and feedback trials. Several participants enunciated words, changed their tone, and broke text into pieces. One participant used the NATO phonetic alphabet to convey words (P12). Participants also exhibited behaviors such as waiting for the transcriptions to stop before speaking, speaking slowly, spelling words out, and speaking louder in order to be properly picked up by the system.

5.3 Theme 3: Acceptance of the System

5.3.1 Subtheme 1: Impact of the Environment and Task. Many participants voiced their concerns with utilizing this system given the scenario. The hesitation of verbally relaying private information outweighed the benefits of the feedback system for some participants, who said that the place and task would determine their opinions on use. P5 explained that they would utilize the feedback system in a normal conversation, during a less invasive task like checking out at the grocery store, or during a private conversation like a doctor's appointment.

5.3.2 Subtheme 2: Effort Utilizing the System. All participants acknowledged that using the feedback system required more effort than that of a normal conversation regardless of control or feedback trials. P3 noted that they were frustrated by having to repeat themselves to the worker and would have liked prior indication to expect this interaction. Some participants noted preferring to use a different tool or system such as a keyboard or handwriting text. Some participants voiced concerns with the time needed to get used to the system and wanted guidelines on how best to approach the scenario.

Despite this acknowledgment of effort, twelve participants were still willing to use the feedback system as they recognized the benefits. However, seven participants said that they preferred not using the system as it required too much effort on their part despite being aware of the overall benefits. Hence, while all participants are aware that the feedback system comes with an increase in cognitive effort and increased improvement in communication accuracy, some of them show a lack of empathy and are unable to accept sharing the cognitive load with the DHH individual.

6 LIMITATIONS AND FUTURE WORK

One limitation of the study was the reliability of our captioning system. There were many times throughout trials where the tooz glasses would pause in the middle of transcriptions and the researcher would have to wait for them to reconnect. These pauses added to participants' frustration during feedback trials, because while the researcher could not see captions, the participant could on their device. The captioning software also was not completely accurate which played a part in user frustration and the need for repetition. Multiple users also pointed out the need for a better UI on the software, particularly mentioning that it does not separate transcriptions based on different speakers, so they were confused to whom the text belonged. The device used for relaying transcriptions to the participant was also a limiting factor. The small screen size resulted in users having to read large blocks of text at one time, making it difficult to keep up with all of the transcriptions.

The fact that our researcher was not a DHH individual was also a limitation. It would be better to have a DHH individual act as the DMV worker in future work as it would increase the validity of the study and ensure that the study is more true to the DHH experience. Furthermore, having a DHH researcher facilitating in the study could open doors to more interesting insights about conversations with hearing individuals and the cognitive load exerted from both parties.

Another limitation was that the task, while good for easily collecting information and prompting conversation, was not the best way to demonstrate the use cases for the feedback system. Some participants questioned the reasoning for why they would utilize the system in a DMV scenario, so this could have affected how they viewed the need for the feedback system. Also, the researcher

was focused on reading captions and writing responses, so they did not look at the participant that much, limiting the chance for participants to engage in natural conversation.

Despite these limitations, the nature of our DMV study was straightforward and left little room for misinterpretation due to the objectiveness of the information, allowing us to collect performance metrics to analyze. In the future, it would be interesting to study the effect of captioning feedback in subjective conversation, where there is no set structure to the task. Conducting the study with a different scenario or task would be a good way to recreate this study in the future and ensure that the system is applicable to others.

Although our participant pool had a diverse range of first and second languages, it would be valuable to explore the effects of different accents, speech behaviors and patterns, and other factors that come with English being a second language. Reproducing the study with improved software, glasses, and UI would be useful to see if participant preferences change based on other variables. Additionally, using speech identification to mask the person using the caption glasses would reduce overlapping captions. Furthermore, future studies could explore how impactful this system is in a group setting. As this study focused on one-on-one conversations, it would be beneficial to know how participants feel using the system when speaking to multiple people.

7 CONCLUSION

The evaluation of captioning with and without speaker feedback in interactions with DHH individuals reveals that captions alone are insufficient for effective communication. Providing feedback to the speaker enhances the interaction by enabling self-correction of mistranscribed text and improving the overall captioning experience. Qualitative data analysis sheds light on varied perspectives, emphasizing concerns about maintaining natural conversation and observed behavioral adjustments. The findings underscore the importance of fostering empathy from the speaker and promoting a deeper understanding of the DHH perspective. This research highlights the need for intentional captioning practices, and calls for increased awareness and knowledge to facilitate more inclusive and effective communication set-ups with DHH individuals, offering valuable insights for future accessibility enhancements.

REFERENCES

[1] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[2] 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*. 1–6.

[3] Gabriel Wade Britain, David Martin, Tyler Kwok, Adam Sumilong, and Thad Starner. 2022. Preferences for Captioning on Emulated Head Worn Displays While in Group Conversation. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers*. 17–22.

[4] Janine Butler, Brian Trager, and Byron Behm. 2019. Exploration of Automatic Speech Recognition for Deaf and Hard of Hearing Students in Higher Education Classes. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 32–42. https://doi.org/10.1145/3308561.3353772

[5] Shelly Chadha, Kaloyan Kamenov, and Alarcos Cieza. 2021. The world report on hearing, 2021. *Bulletin of the World Health Organization* 99, 4 (April 2021), 242–242A. https://doi.org/10.2471/BLT.21.285643

[6] Rucha Deshpande, Tayfun Tuna, Jaspal Subhlok, and Lecia Barker. 2014. A crowdsourcing caption editor for educational videos. In *2014 ieee frontiers in education conference (fie) proceedings*. IEEE, 1–8.

[7] Harvey Dillon. 2012. *Hearing aids* (2nd ed.). Thieme Medical Publishers, United States.

[8] Peter Feng, David Martin, and Thad Starner. 2023. ToozKit: System for Experimenting with Captions on a Head-worn Display. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*. 175–178.

[9] Gyeong-Nam Gimhae. 2013. Six human factors to acceptability of wearable computers. *International Journal of Multimedia and Ubiquitous Engineering* 8, 3 (2013), 103–114.

[10] Google. 2019. Live Transcribe | Speech to Text App | Android. https://www.android.com/accessibility/live-transcribe/.

[11] Richard S. Hallam and Roslyn Corney. 2014. Conversation tactics in persons with normal hearing and hearing-impairment. *International Journal of Audiology* 53, 3 (March 2014), 174–181. https://doi.org/10.3109/14992027.2013.852256 Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/14992027.2013.852256.

[12] Dhruv Jain, Rachel Franz, Leah Findlater, Jackson Cannon, Raja Kushalnagar, and Jon Froehlich. 2018. Towards Accessible Conversations in a Mobile Context for People Who Are Deaf and Hard of Hearing. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 81–92. https://doi.org/10.1145/3234695.3236362 event-place: Galway, Ireland.

[13] 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. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300324 event-place: Glasgow, Scotland Uk.

[14] Anna Marie Jilla, Carole E Johnson, and Nick Huntington-Klein. 2023. Hearing aid affordability in the United States. *Disability and Rehabilitation: Assistive Technology* 18, 3 (2023), 246–252.

[15] Ashley Miller, Joan Malasig, Brenda Castro, Vicki L. Hanson, Hugo Nicolau, and Alessandra Brandão. 2017. The Use of Smart Glasses for Lecture Comprehension by Deaf and Hard of Hearing Students. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 1909–1915. https://doi.org/10.1145/3027063.3053117

[16] Anant Mittal, Meghna Gupta, Roshni Poddar, Tarini Naik, Seethalakshmi Kuppuraj, James Fogarty, Pratyush Kumar, and Mohit Jain. 2023. Jod: Examining Design and Implementation of a Videoconferencing Platform for Mixed Hearing Groups. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 43, 18 pages. https://doi.org/10.1145/3597638.3608382

[17] Alex Olwal, Kevin Balke, Dmitrii Votintcev, Thad Starner, Paula Conn, Bonnie Chinh, and Benoit Corda. 2020. Wearable Subtitles: Augmenting Spoken Communication with Lightweight Eyewear for All-day Captioning. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1108–1120. https://doi.org/10.1145/3379337.3415817

[18] Phil Parette and Marcia Scherer. 2004. Assistive technology use and stigma. *Education and training in developmental disabilities* 39, 3 (2004), 217–226.

[19] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3173867

[20] Halley Profita, Reem Albaghli, Leah Findlater, Paul Jaeger, and Shaun K. Kane. 2016. The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4884–4895. https://doi.org/10.1145/2858036.2858130

[21] T.E. Starner. 2003. The enigmatic display. *IEEE Pervasive Computing* 2, 1 (2003), 15–18. https://doi.org/10.1109/MPRV.2003.1186720

[22] Tooz. 2023. Tooz DevKit. https://tooz.com/devkit/.

[23] M. Wald. 2011. Crowdsourcing correction of speech recognition captioning errors. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility* (Hyderabad, Andhra Pradesh, India) *(W4A '11)*. Association for Computing Machinery, New York, NY, USA, Article 22, 2 pages. https://doi.org/10.1145/1969289.1969318

A PARTICIPANT DEMOGRAPHICS

Participant ID	Age	Gender	Languages Known
P1	20	Female	Mandarin, English
P2	26	Female	Hindi, English
P3	19	Female	English, American Sign Language
P4	21	Male	English
P5	30	Female	English
P6	21	Male	English
P7	23	Female	Telugu, Hindi, English
P8	21	Male	English
P9	22	Male	English
P10	20	Male	English, Korean, Chinese
P11	24	Female	English, German
P12	20	Male	English, Chinese
P13	20	Female	Chinese, English
P14	20	Male	English, Chinese
P15	20	Female	English, Mandarin
P16	20	Female	Russian, English
P17	20	Male	English, Korean
P18	21	Female	English, Korean
P19	20	Male	English, Korean
P20	20	Male	English

Table 1: Participant Demographics (Languages in order of proficiency)

B SEMI-STRUCTURED INTERVIEW QUESTIONS

(1) Describe your experience speaking with the DMV worker.
(2) Describe your experience viewing the captions and without viewing them during the conversation with the DMV worker.
(3) Did seeing your transcription affect how you communicated? If so, how?
(4) What would have made this experience better?
(5) Did you have a preference for one method over the other?
(6) Is there anything else you want to share?

C SUS SCORES, DURATION TAKEN, MISTRANSCRIPTIONS, AND CORRECTIONS: CONTROL

Participant ID	Time Taken	Number of Mistakes	SUS	Prompted Corrections	Unprompted Corrections
P1	11:29:00	4	77.59	12	1
P2	8:01:00	4	82.76	11	0
P3	8:42:00	8	84.48	9	0
P4	6:47:00	4	55.17	5	0
P5	6:31:00	11	77.59	6	0
P6	5:45:00	0	84.48	7	0
P7	6:42:00	3	74.14	10	2
P8	5:31:00	5	86.21	5	0
P9	7:40:00	4	89.66	9	0
P10	7:52:00	4	63.79	7	0
P11	6:05:00	3	74.14	3	0
P12	6:30:00	1	94.83	3	0
P13	5:37:00	2	100	2	0
P14	5:48:00	0	96.55	3	0
P15	6:34:00	12	77.59	10	0
P16	4:58:00	2	84.48	5	0
P17	5:53:00	2	67.24	6	0
P18	5:24:00	0	98.28	2	0
P19	5:29:00	1	98.28	4	0
P20	5:32:00	2	68.97	6	0
μ	**6:06:47**	**2.93**	**83.87**	**5.36**	**0.14**
σ	**0:51:17**	**3.00**	**12.79**	**2.79**	**0.53**

D NASA TLX SCORES: CONTROL

Participant ID	Mental	Physical	Temporal	Performance	Effort	Frustration	Overall
P1	0.00	0.00	0.00	100.00	0.00	0.00	**16.67**
P2	30.00	10.00	25.00	75.00	25.00	35.00	**33.33**
P3	20.00	0.00	10.00	90.00	30.00	60.00	**35.00**
P4	60.00	50.00	25.00	90.00	65.00	55.00	**57.50**
P5	15.00	5.00	5.00	90.00	10.00	10.00	**22.50**
P6	15.00	0.00	5.00	95.00	5.00	25.00	**24.17**
P7	70.00	10.00	45.00	95.00	65.00	20.00	**50.83**
P8	15.00	0.00	20.00	85.00	15.00	5.00	**23.33**
P9	0.00	0.00	15.00	80.00	10.00	10.00	**19.17**
P10	80.00	0.00	70.00	100.00	75.00	65.00	**65.00**
P11	0.00	0.00	0.00	100.00	10.00	0.00	**18.33**
P12	50.00	0.00	30.00	85.00	50.00	35.00	**41.67**
P13	0.00	0.00	0.00	100.00	0.00	0.00	**16.67**
P14	0.00	0.00	0.00	95.00	10.00	0.00	**17.50**
P15	0.00	0.00	0.00	85.00	10.00	0.00	**15.83**
P16	0.00	0.00	0.00	80.00	10.00	0.00	**15.00**
P17	40.00	5.00	45.00	65.00	45.00	55.00	**42.50**
P18	55.00	0.00	25.00	20.00	55.00	35.00	**31.67**
P19	0.00	0.00	0.00	85.00	5.00	0.00	**15.00**
P20	15.00	0.00	15.00	85.00	60.00	15.00	**31.67**
μ	**23.21**	**1.07**	**18.93**	**82.86**	**30.00**	**17.14**	**28.87**
σ	**29.52**	**2.89**	**22.03**	**20.54**	**26.53**	**22.08**	**15.74**

Table 2: NASA TLX scores across all participants for control condition

E SUS SCORES, DURATION TAKEN, MISTRANSCRIPTIONS, AND CORRECTIONS: FEEDBACK

Participant ID	Time Taken	Number of Mistakes	SUS	Prompted Corrections	Unprompted Corrections
P1	11:10:00	0	75.86	9	3
P2	10:54:00	1	84.48	10	4
P3	7:52:00	2	41.38	8	0
P4	8:39:00	2	46.55	6	1
P5	8:30:00	7	75.86	6	2
P6	6:52:00	1	87.93	6	1
P7	8:12:00	3	77.59	4	7
P8	6:42:00	10	81.03	7	2
P9	7:00:00	2	89.66	4	7
P10	8:17:00	1	68.97	2	4
P11	6:27:00	1	68.97	2	3
P12	13:07:00	0	82.76	2	7
P13	5:46:00	0	63.79	2	1
P14	6:18:00	0	100	1	7
P15	7:02:00	12	91.38	3	7
P16	5:39:00	4	72.41	6	0
P17	5:10:00	0	86.21	1	3
P18	5:18:00	2	72.41	2	0
P19	6:04:00	0	89.66	4	2
P20	8:38:00	1	96.55	3	5
μ	**7:07:09**	**2.57**	**81.53**	**3.07**	**3.93**
σ	**2:02:39**	**3.80**	**11.20**	**1.77**	**2.73**

Table 3

F NASA TLX SCORES: FEEDBACK

Participant ID	Mental	Physical	Temporal	Performance	Effort	Frustration	Overall
P1	0.00	0.00	0.00	100.00	0.00	0.00	**16.67**
P2	75.00	5.00	20.00	70.00	45.00	10.00	**37.50**
P3	60.00	10.00	20.00	40.00	70.00	60.00	**43.33**
P4	60.00	0.00	25.00	70.00	55.00	65.00	**45.83**
P5	5.00	5.00	5.00	90.00	5.00	25.00	**22.50**
P6	25.00	0.00	5.00	85.00	10.00	25.00	**25.00**
P7	85.00	25.00	65.00	95.00	65.00	15.00	**58.33**
P8	35.00	0.00	25.00	70.00	80.00	20.00	**38.33**
P9	0.00	0.00	5.00	80.00	10.00	5.00	**16.67**
P10	75.00	0.00	75.00	100.00	75.00	60.00	**64.17**
P11	25.00	0.00	0.00	80.00	20.00	10.00	**22.50**
P12	60.00	0.00	30.00	65.00	35.00	55.00	**40.83**
P13	0.00	0.00	0.00	100.00	10.00	0.00	**18.33**
P14	5.00	0.00	0.00	85.00	15.00	0.00	**17.50**
P15	0.00	0.00	0.00	85.00	15.00	0.00	**16.67**
P16	0.00	0.00	0.00	80.00	30.00	10.00	**20.00**
P17	45.00	5.00	40.00	90.00	55.00	55.00	**48.33**
P18	55.00	50.00	45.00	40.00	35.00	30.00	**42.50**
P19	0.00	100.00	95.00	75.00	85.00	80.00	**72.50**
P20	70.00	65.00	60.00	60.00	50.00	45.00	**58.33**
μ	**32.50**	**17.50**	**31.43**	**78.93**	**41.43**	**27.50**	**38.21**
σ	**32.21**	**31.73**	**32.61**	**16.43**	**26.85**	**26.66**	**19.84**

Table 4: NASA TLX scores across all participants for feedback condition