# Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One Meetings

Larwan Berke
Rochester Institute of Technology
152 Lomb Memorial Drive
Rochester, NY 14618 USA
larwan.berke@mail.rit.edu

Christopher Caulfield
Rochester Institute of Technology
152 Lomb Memorial Drive
Rochester, NY 14618 USA
cxc4115@rit.edu

Matt Huenerfauth
Rochester Institute of Technology
152 Lomb Memorial Drive
Rochester, NY 14618 USA
matt.huenerfauth@rit.edu

## ABSTRACT
Recent advances in Automatic Speech Recognition (ASR) have made this technology a potential solution for transcribing audio input in real-time for people who are Deaf or Hard of Hearing (DHH). However, ASR is imperfect; users must cope with errors in the output. While some prior research has studied ASR-generated transcriptions to provide captions for DHH people, there has not been a systematic study of how to best present captions that may include errors from ASR software nor how to make use of the ASR system's word-level confidence. We conducted two studies, with 21 and 107 DHH participants, to compare various methods of visually presenting the ASR output with certainty values. Participants answered subjective preference questions and provided feedback on how ASR captioning could be used with confidence display markup. Users preferred captioning styles with which they were already most familiar (that did not display confidence information), and they were concerned about the accuracy of ASR systems. While they expressed interest in systems that display word confidence during captions, they were concerned that text appearance changes may be distracting. The findings of this study should be useful for researchers and companies developing automated captioning systems for DHH users.

## CCS Concepts
• **Human-centered computing~Empirical studies in accessibility** • **Human-centered computing~Accessibility design and evaluation methods** • *Social and professional topics~Assistive technologies*

## Keywords
Deaf and Hard of Hearing; Automatic Speech Recognition; Real-Time Captions; Communication; User Study; Feedback.

## 1. INTRODUCTION
Advances in speech and language technology can benefit people who are Deaf or Hard of Hearing (DHH) by providing access to spoken information. Many of these users currently benefit from a wide range of services such as e-mail/instant messaging [3], American Sign Language (ASL) in-person interpretation, real-time transcription or captioning services, and Video Relay Service [6]. However, in many work or education settings, DHH individuals may lack access to sign language interpreting services due to their high cost or the need to arrange such services in advance. Further, 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, rather than sign language interpretation.

With recent advances in Automatic Speech Recognition (ASR), researchers have considered whether fully automatic solutions for providing text transcriptions of spoken language could be useful for DHH users [35]. An exploratory survey of DHH participants by Elliot *et al.* [10] revealed that users were agreeable to the idea of having ASR support their workplace conversations, but the study did not provide users with an opportunity to try a prototype of this technology. Kawas *et al.* [23] found that participants felt more autonomous when using a current ASR platform for real-time captioning, but their study focused on a classroom setting, and with a small number of participants.

While these prior studies suggest the promise of ASR for workplace meeting captioning, an empirical study with a large number (100+) of DHH participants is needed to identify factors important to users. Similar to [23], we expect that DHH users may be better able to provide feedback after they have an opportunity to try a prototype ASR captioning system, in our case, with a simulated workplace meeting, and with word-confidence information displayed. This paper describes a pair of studies we have conducted with 128 total DHH users trying such a prototype and providing feedback.

### 1.1 ASR Tools for DHH Individuals
Several researchers have investigated how to produce captioning tools to benefit DHH users via ASR technologies. Some focused on non-real-time applications, e.g., ASR producing a transcript of classroom lectures, after the lecture is concluded, to enable students to review material after class [44]. Others examined whether ASR can automatically caption the content of online videos [38]. In these above settings, the user must wait for offline ASR processing.

Other researchers have investigated semi-automated real-time applications of ASR for DHH users: In [24], students viewed lectures with real-time captioning, to achieve good accuracy, their ASR used a human-prepared dictionary for each course. As another example, in some Communication Access Realtime Translation (CART) services, a professional clearly re-speaks the words of the lecturer into a high-end microphone, to achieve high ASR accuracy [29]. While ASR technology has improved, there are still errors in the output, especially in noisy and complex environments. To boost the accuracy of imperfect ASR, some researchers have created systems in which human overseers fix mistakes in ASR output [15] or have crowdsourced the task of transcribing audio (thereby bypassing the use of ASR) [17]. The business model for these services requires some regular payment for the human labor.



*ASSETS'17, October 29-November 1, 2017, Baltimore, MD, USA*

ACM ISBN 978-1-4503-4926-0/17/10…$15.00
https://doi.org/10.1145/3132525.3132541

Some researchers have experimented with fully automated real-time systems for DHH users: Some examined whether augmented reality glasses can overlay ASR-produced captions on the field of view of DHH users [31], but they found that the quality of several ASR engines was insufficient for DHH users, due to noisy environments in daily life that led to low accuracy, e.g. 60%. Kawas *et al.* [23] performed a qualitative study of ASR in classrooms and had similar findings; their participants struggled with low quality captioning from a real-time ASR application.

In contrast to prior work, our research focuses on a context in which ASR may work better: during live one-on-one meetings between a DHH individual and a hearing person. By removing a major source of noise from the environment (e.g. classroom with crowds and background sounds) and focusing on one speaker, ASR engines are capable of producing output with higher accuracy. Furthermore, contrast this setting to a public lecture: A lecturer may not change how they speak based on ASR output since they are busy attending to their entire diverse audience; in contrast, during a live one-on-one meeting, a speaker might adjust their voice or speaking style to lead to better ASR results if they notice errors in the ASR output.

We focus on a scenario in which ASR results are displayed on a mobile tablet device viewable by the DHH participant in a one-on-one meeting with hearing individuals. To prototype how captions may appear, we produced simulation videos of a business meeting (Figure 1) with a camera focused on an actor who is sitting at a table (across from the camera). The layout for this mock conversation is shown in Figure 1 with the text captions appearing in a black area at the bottom of the video window, containing the output from the ASR engine, representing where the DHH user might view a tablet device below the line of-sight with their conversational partner. Additional details about these videos appear in Section 1.3.


**Figure 1. The prototype tool examined in this work**

## 1.2 Displaying Confidence

Prior research has found that users may benefit from captions, even if only a portion of the speech audio was successfully captioned. For example, a participant in one study using ASR captioning said: "knowing the context and searching for keywords are essential steps to build their capacity of understanding." [34] This finding is valuable since ASR systems are currently not perfectly accurate in identifying the words spoken, especially in noisy environments.

There is additional information we can provide to DHH users: ASR engines assign confidence metrics to the words they hypothesize are being spoken in an audio input; this information could be utilized by the viewer to know which portions of the text they should trust. Most ASR applications do not convey this information to the user. We investigate whether knowing the difference between high-confidence (confident) and low-confidence (uncertain) words, given some threshold, may benefit DHH users of automatically produced captions (which will naturally contain errors).

In addition to investigating *whether* providing ASR confidence information benefits DHH users, we also examine *how* we should display this information visually. We must consider how to reduce the complexity of the information that is presented to users. In a live one-on-one meeting, the DHH participant must read the caption and attend to the face and body language of their conversational partner. We seek a method of displaying uncertain words/phrases in a way that does not create additional communication barriers for DHH users by making the visual presentation too complex.

Several researchers have studied how to present ASR captions on screen, including whether to indicate if the ASR was not confident about some of the words. Figure 2 lists how prior researchers have modified the appearance of captions, including a few researchers who have specifically examined how to possibly convey the ASR confidence level to viewers (which is the focus of our research).

**Changing the appearance to convey ASR confidence:**

| Appearance Changed | Prior Published Research |
|---|---|
| Font Change | Piquard-Kipffer *et al.* [34] |
| Font Color | Shiver and Wolfe [39] |
| Underlining | Vertanen and Kristensson [42] |

**Changing the caption appearance for other reasons:**

| Appearance Changed | Prior Published Research |
|---|---|
| Colored Borders | Vy [43] |
| Dynamic Positioning | Hong *et al.* [18] |
| Dynamic Size | Wang *et al.* [45] |
| Emoji | Lee *et al.* [28] |
| Removal of Text | Ferdiansyah and Nakagawa [12] |
| Syllables | Secară [38] |
| Text Spacing | Nambo *et al.* [32] |
| Tracked Display | Kushalnagar *et al.* [25] |
| Transparency | Miller *et al.* [30] |

**Figure 2. Captioning Display Styles**

Piquard-Kipffer's team [34] indicated words for which the ASR was confident using **bold** text, and uncertain words, in regular text. They experimented with conveying ASR-uncertain words with phonetic spelling (like a dictionary pronunciation listing), but DHH users did not like that approach. Noting that only those participants who knew phonology of the spoken language could benefit from the extra information, the authors explained that their method may not benefit members of the DHH community who utilize sign language as their primary language.

Shiver and Wolfe [39] used white text (on a black background) to indicate words for which the ASR was confident, and darker gray text for the word for which ASR was uncertain. Some participants in their study liked this approach. However, when the authors tried to measure the impact of this addition through comprehension-question testing, they did not measure an effect. The authors used caution in drawing conclusions from their comprehension-question results: "even when there were no captions available, participants answered more than 50% of the content questions correctly... Also, the performance on the perfectly-captioned video was lower than that on the two videos captioned through ASR." [39]

DHH users were not the focus of [42], but those researchers studied the use of red underlines to indicate ASR-uncertain words in the text output of a dictation application. They studied how visualizing ASR errors could help hearing individuals correct the output, but the authors noted that confidence visualization did not improve participants' ability to find errors [42]. However, the authors noted

that when their system accurately added "uncertain" underlines to erroneous words, participants were more successful at fixing the ASR dictation output, as compared to a baseline condition with no underlining. Their work revealed that the potential benefits of uncertainty marking may depend on the ASR accuracy level.

While prior work has examined methods of marking words in captions, e.g. [34, 39], the literature lacks an empirical comparison of various approaches for conveying confidence, based on reactions from DHH users – especially in the context of one-on-one meetings between a DHH individual and a hearing person. Such a study can identify promising design directions, which could be further explored and refined through additional design work and evaluation.

## 1.3 Prototype and Stimuli for Our Studies

To support our study, we built a prototype that uses ASR to process speech from videos to automatically add captions, with appearance dependent on the confidence numbers in the ASR output. Figure 1 showed an example of the output of our current prototype.

As discussed in Section 1.1, while we envision a future application in which users may hold a tablet device that displays captions during a one-on-one meeting with a hearing individual, we have created videos of a mock business meeting scenario with captions displayed onscreen. Crabb *et al.* [7] recommended the captioning be positioned below and outside of a video frame when captioning online videos. In our case, we are simulating where a person may hold a tablet device displaying live captions, below the line of sight of their conversational partner; in the prototype videos, the user can see the real-world environment to the left and the right of the captions, as they would to the left and right of a tablet they may hold at that location in real life. The ASR output appears onscreen one word at a time, similar to C-Print [11], rather than appearing as entire phrases or sentences at once. Further, onscreen text is limited to 3 lines so that it does not take up too much visual space, following recommendations in Kushalnagar *et al.* [26].

In Figure 1, red indicated words for which the ASR engine had a low confidence value. The speaker actually said "which college *career fairs they will* be attending based on *the candidate requirements.*" The prototype can be configured to format text in a variety of ways, and the automatically generated caption text can be modified manually to insert or remove errors, for testing purposes. Determining the optimal threshold of ASR confidence for determining when to apply text formatting is an open question because it greatly depends on the specific ASR engine. Another open design question is the span of the text to use as the granularity for the uncertainty marking. For the research in this paper, we shall assume a per-word confidence based on ASR engine confidence rather than calculating confidence over multi-word chunks of text.

In this paper, we use this prototype to compare various methods of displaying word confidence to DHH users and elicit their feedback. With Figure 2 as a starting point, we wanted to select a set of visual parameters to convey word-confidence to a DHH user reading captions. To reduce our design space, we considered captioning standards, e.g. digital television captioning includes basic ASCII plus a few special characters [2], and some standards recommend the look and speed of video captions [9]. Other researchers recommend avoiding animated text [36], avoiding small/large font sizes [5], avoiding text/background color combinations that clash or are too similar [8], and encouraged the use of sans serif fonts for legibility [21]. Figure 3 displays the caption appearance options we decided to investigate in a preliminary study (Section 2.1). We included some "opposites" such as *Bold on Confident* (bold_c) and *Bold on Uncertain* (bold_u), but some caption markup styles lacked a logical opposite, e.g. *Empty Underline on Uncertain* (del_u).

To produce stimuli that could be repeated across participants with variations in the captioning display, we simulated a one-on-one meeting by creating a short film of a business meeting. We wrote a script with someone discussing upcoming plans for a project with the person viewing the video: A human resources office was recruiting job candidates at an upcoming event. An actor performed the meeting script in a sound-proof video studio, with an office desk and plain background (Figure 1). The actor's voice was captured by a professional microphone on an overhead boom. The script was composed of 12 stimulus paragraphs, with an average of 88.3 words in each paragraph (ranging in duration from 19 to 46 seconds).

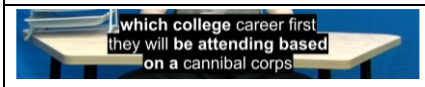
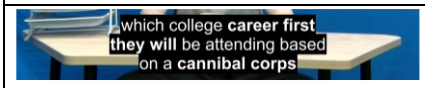
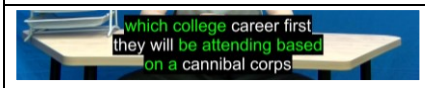
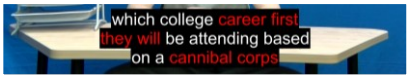
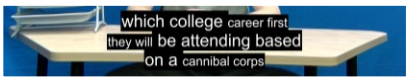
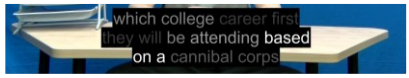
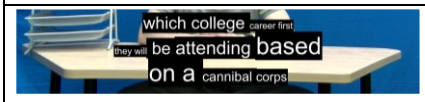
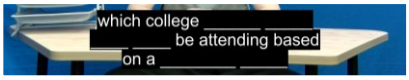
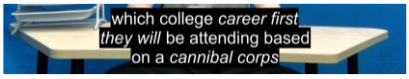
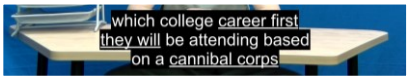
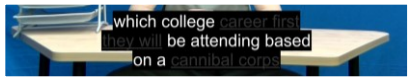
| Example of Markup on Captioning | Description (label) |
|---|---|
|  | Baseline condition: no markup (no_change) |
|  | Bold on Confident (bold_c) |
|  | Bold on Uncertain (bold_u) |
|  | Green on Confident (color_c) |
|  | Red on Uncertain (color_u) |
|  | Small font size on Uncertain (size_u) |
|  | Levels of gray color based on confidence (r_gray) |
|  | Levels of font size based on confidence (r_size) |
|  | Empty underline on Uncertain (del_u) |
|  | Italics on Uncertain (it_u) |
|  | Underline on Uncertain (ul_u) |
|  | Underline and gray color on Uncertain (ul_gray_u) |

**Figure 3. Markup conditions in the pilot study (Section 2.1)**

To generate realistic ASR output (containing errors), the audio from the videos was processed using the Sphinx open-source ASR engine [27]. Accuracy is reported using Word Error Rate (WER), based on the number of substitutions, insertions, and deletions in the output when compared to the actual script. The average WER of our stimuli was 23.2%. We selected this rate because we needed sufficient errors to appear in our caption texts so that participants would see at least a few words with markup in each stimulus. If the WER were any worse, we were concerned that participants would find the captions so unhelpful that they would not remain engaged during the study.

To decide which words should receive markup, a confidence threshold must be defined so that words can be categorized as either *Confident* (above the threshold) or *Uncertain* (below it). Based on our observations and feedback from a small number of DHH users who viewed initial prototype videos, we selected a probability threshold of 0.995. (It was not the intent of this study to determine an optimum threshold value; we leave this for future work.) It is important to note that the confidence value is not a perfect indication of the ASR accuracy for individual words. That is, some words correctly recognized by the ASR may be labeled as "uncertain," and conversely, some words that were incorrectly recognized by the ASR may be labeled as "confident." For the 1060 words in our stimuli, 789 were correct and confident, 197 were correct yet uncertain, 39 were incorrect yet confident, and 69 were incorrect and uncertain.

Finally, all 12 stimulus-paragraph videos were generated using all 12 markup conditions in Figure 3, to yield 144 videos.

# 2. RESEARCH QUESTIONS & METHODS

This paper investigates several research questions:

**RQ1:** Are DHH users receptive to the idea of using ASR to caption spoken content during one-on-one meetings?

**RQ2:** What issues/factors do DHH users consider important when they are using ASR captioning as an assistive technology?

**RQ3:** When viewing captions of speech in a one-on-one meeting, what confidence markup do DHH users subjectively prefer?

**RQ4:** What applications do DHH users believe are most suitable for the use of ASR-based automatic captioning tools?

As we mentioned in Section 1, prior researchers have found that displaying caption markup for ASR applications has promise: in particular, as measured through subjective preferences of users. However, we did not find prior work that focused on one-on-one meetings nor determined which method of visual markup is preferred by DHH users. One study that closely resembled our research questions was done by Kawas *et al.*, however they did not exclusively focus on ASR as a captioning tool for one-on-one meetings nor did they discuss displaying confidence [23]. While prior guidelines for caption display helped to narrow our design space, we still had quite a long list of word-markup parameters to consider (Figure 2). It was infeasible to expect our participants to give reasonable answers when presented with this many choices. Therefore, we split our research into two parts: a "pilot" study to narrow down the markup choices, and a larger follow-up study. The design of the first (pilot) study is a single-question quantitative survey, while the second, larger "comparison" study used both quantitative and qualitative methods to analyze responses to subjective scalar questions and open-ended feedback questions.

## 2.1 Pilot Study to Narrow the Markup Set

The goal of this pilot study was to identify a subset of the most promising visual markup styles. DHH participants viewed videos of a mock one-on-one meeting situation, with captions displayed. This within-subjects study had 12 conditions: the 11 visual display styles listed in Figure 3 and a baseline condition (white text on black background, with no indication of word confidence). Each participant viewed and subjectively evaluated a total of 12 videos, with one video of each of the different markup methods. The 12 conditions were assigned to the 12 paragraph-length stimuli videos (Section 1.3) using a Latin-squares schedule, to rotate conditions among the individual videos. The 12 video stimuli of the business meeting were presented to participants in chronological order.

We designed videos on a MacBook Pro 15-inch laptop with a web browser displayed stimuli and questions; the laptop was pre-installed with a PHP server and hosted the stimuli locally. A native ASL signer research assistant was with the participant at all times in order to provide guidance on using the application and technical troubleshooting when necessary. The application presented a typical HTML form that the participant completed and pressed "Next" to proceed through the study.

We reached out to participants by e-mail and flyers on the university campus. Participants were eligible if they answered "yes" to both screening questions: Are you 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 50-minute study. A total of 21 DHH individuals participated, with self-identified gender of 13 males and 8 females and self-identified hearing-status of 14 people who are Deaf and 7 people who are Hard of Hearing.

After completing consent forms and a demographic questionnaire, participants were shown an introductory video to introduce the scenario and explain the purpose of the visual markup. Participants were informed that the words they would see were produced by a computer that was trying to identify what was spoken automatically; further, they were informed that the computer wasn't always confident in what it heard, and the changes in appearance of the text indicated this. After each video stimulus, participants answered a Yes/No question: "Did you like the style of captioning in this video clip?" The question was modeled after the Quality of Perception (*QoP*) scale shown to be effective for DHH individuals in [16]. Figure 4 displays mean scores (No=0/Yes=1) for each markup style. The differences shown in Figure 4 are not statistically significant; it was not the intent of this pilot study to conclusively select a best markup method: Our goal was to identify the most promising candidates from this large set of visual mark-up styles, which would be further examined in a large follow-up study. Section 2.2.1 discuss how this selection was made.

At the end of the study, participants were asked to write open-ended feedback comments about any aspect of the study.

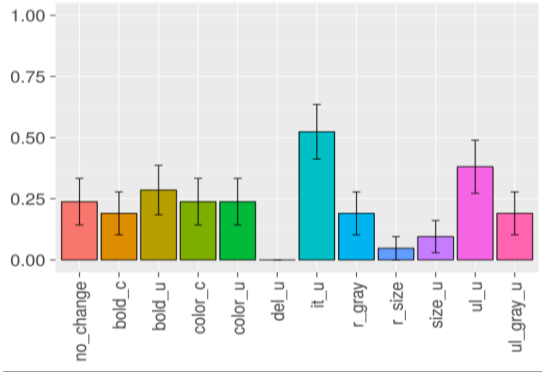



**Figure 4. Pilot Study Preference Results**

## 2.2 Larger Study

A subset of the most preferred visual markup styles in the preliminary study were compared in a subsequent study containing a larger number of participants. While the design of this larger study was mostly identical to the first study, a few modifications and refinements were made, summarized in the following subsections.

### *2.2.1 New List of Markup Styles*

The first step was to reduce our set of markup styles by analyzing the results from the preliminary study. Given practical limits on the number of DHH participants who could be recruited for this study, we wanted to reduce our study design to only four conditions: a baseline and three visual markup styles. We used the following criteria to select the four conditions in this follow-up study:

1. *Retain the baseline:* The captions without any visual markup (no_change) were retained for the larger, comparison study so that we could compare this ubiquitous form of captioning to the new captioning methods that included visual markup.
2. *Retain the markup styles that received the highest subjective preference scores in the preliminary study:* Videos that used *italics* to indicate uncertain words (it_u) had the highest scores, and videos that used underlining to indicate uncertain words (ul_u) had the second highest scores in Figure 4.
3. *Eliminate styles that markup confident words:* The subjective preference results were less clear-cut for the remaining videos, but videos that utilized color or **boldness** seemed to have the highest scores among those that remain. In feedback comments, several participants indicated that applying markup to confident words was visually distracting. Since this issue will only be exacerbated as the accuracy of ASR systems improves over time, we decided to eliminate any markup styles that apply visual markup to text that was confidently identified by the ASR system (bold_c, color_c). Therefore, we were left to consider whether to retain the Bold on Uncertain (bold_u) or Color on Uncertain (color_u) markup style for inclusion in our larger, comparison study.
4. *Considering colorblindness accessibility:* The set of visual markup styles in the preliminary study included both Color on Uncertain (red) and Color on Confident (green); with the red/green meant to convey a negative/positive distinction. However, one participant in our preliminary study indicated that he had difficulty distinguishing the red/green colors from the "unmarked" white text. For this reason, we decided to merge the Bold on Uncertain and Color on Uncertain styles to create a new style, "Bold and Color on Uncertain," that redundantly conveyed the information through both color and boldness. While the specific way in which an individual may experience color vision deficiency may be unique [22, 13], there are trends: We selected a yellow color (red 100%, green 100%, blue 0%) which retained contrast with both the black background and also with the unmarked white text, across a variety of common forms of colorblindness.

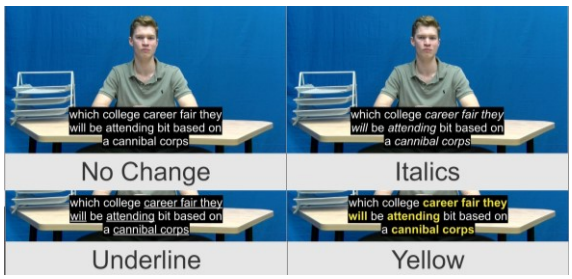



**Figure 5. Markup conditions in larger study (Section 2.2)**

Therefore, we settled on a list of 4 markup styles for the expanded study: *Baseline ("No Change"), Italics on Uncertain, Underline on Uncertain*, and *Yellow+Bold on Uncertain* – as shown in Figure 5.

### *2.2.2 Comprehension Questions*

After each video clip, we asked participants four multiple-choice comprehension questions about information conveyed in the speech audio of the stimuli videos. We used "text-explicit" (TE) style comprehension questions [20] wherein the participant is not required to infer from information in memory. None of our research questions specifically focus on whether different caption markup may lead to different comprehension score results: Therefore, our rationale for including these questions was simply to introduce pressure on the participants to understand what was said, to more realistically simulate using captioning in live meetings, so that our participants would have a higher fidelity prototype experience, prior to answering preference questions at the end of the study.

### *2.2.3 Preference Questions*

Since we reduced the number of markup styles in this larger study, there was time to add questions. Yannakakis and Hallam [47] found that participants may provide different subjective evaluations when they are asked to provide a subjective rating for an item in isolation or when they are asked to rank a set of items; their work suggested the merit of utilizing both methods, especially when there may be serial-presentation contrast effects in a study design, despite the use of randomized Latin squares ordering for the markup styles. In addition to asking subjective preference questions after each markup type (including 5-point Likert-type items and binary Yes/No questions), after all 12 videos were shown, participants in this larger study were asked to rank the 4 captioning styles. The participants were presented a page with pictures of all 4 styles to remind them of each style, and they used four drop-down lists with the markup styles to provide a ranking.

Finally, at the end of the study we asked participants a battery of open-ended questions to elicit their opinion on the confidence markup and their opinions on the use of ASR for captioning. Our questionnaire and videos can be downloaded from our website: http://latlab.ist.rit.edu/assets2017dhh

### *2.2.4 Number of Participants and Demographics*

We recruited a total of 107 DHH participants (those that participated in our pilot study were excluded) from our university campus and they self-identified their hearing status as (69 Deaf, 36 Hard-of-Hearing, 2 other), and gender as (59 males and 48 females). Participants' ages ranged from 18-30 years old.

# 3. QUANTITATIVE RESULTS

This section presents the results of our quantitative analysis of the closed-ended question items during our larger study.

After each video, we asked participants the same question in the pilot study, "Did you like this style of captioning?" Figure 6 displays the mean scores (No=0/Yes=1) for each markup style. A Kruskal-Wallis test indicated significance for the factor of markup style [$\chi 2$=17.587, DF=3, $p$=0.000535]. To compare the markups in a pairwise manner, a Wilcoxon rank-sum test with Bonferroni correction was used for post-hoc analysis; the following pairs differed significantly in their medians: *No Change* / *Underline*: $p$= 0.00085, *No Change* / *Yellow*: $p$=0.00211, *No Change* / *Italics:* $p$=0.02307. Notably, our participants had significantly higher preference scores for the markup condition with *No Change*, in which no confidence information was displayed.

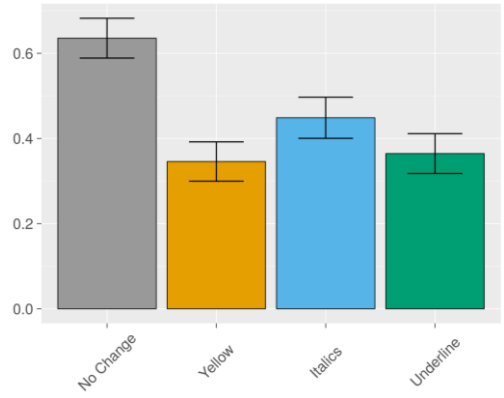



**Figure 6. Larger Study: Preference Responses (binary)**

We followed up on this question with another binary question: "Was this style of captioning distracting?" Figure 7 displays the mean scores (No=0/Yes=1) for each markup style. The Kruskal-Wallis test indicated significance [$\chi2$=22.692, DF=3, $p$=4.682e-05]. Post-hoc pairwise Wilcoxon signed-rank tests revealed significant differences for the following pairs: *No Change / Yellow*: $p$=3.7e-05, *No Change / Underline*: $p$=0.00035, *Italics / Yellow*: $p$=0.00187. Participants seemed to prefer captions without confidence markup or which used italics to convey confidence, from the perspective of reducing distraction.

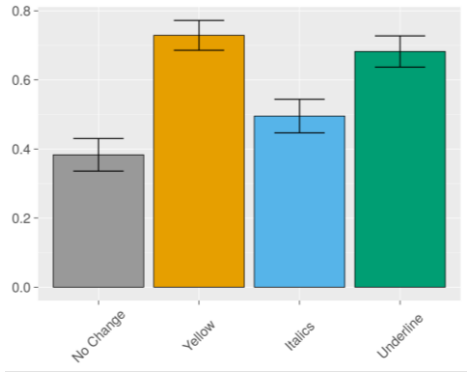



**Figure 7. Larger Study: Distracting Responses (binary)**

The participants were asked a 5-point Likert question, "I think this captioning style would be helpful during face-to-face meetings with hearing people." Figure 8 displays responses for each markup style. The Kruskal-Wallis test did not indicate significance for this question [$\chi2$=3.771, DF=3, $p$=0.2873].

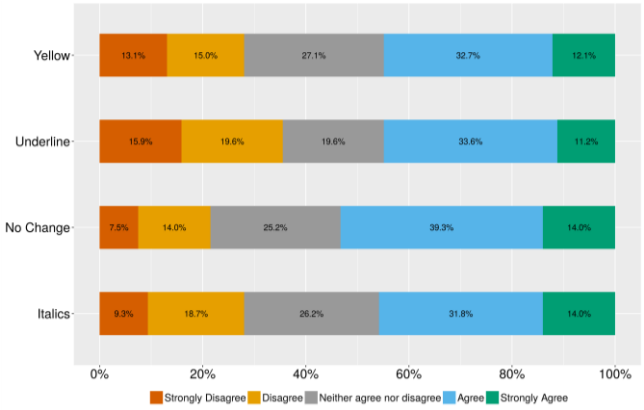



**Figure 8. Larger Study: Helpful Responses (Likert)**

Lastly, we asked participants to rank the markup styles at the end of the experiment with the question "Please select and rank the captioning display styles." Figure 9 displays the percentages of the ranks for each markup style. The Friedman rank sum test indicated significance [$\chi2$=26.559, DF=3, $p$=7.284e-06]. Post-hoc pairwise Wilcoxon signed-ranks test with Bonferroni corrections revealed significant differences for the following pairs: *No Change / Underline*: $p$=0.00014, *No Change / Yellow*: $p$=0.00025, *No Change / Italics*: $p$=0.01576. Similar to the results on the binary preference questions (Figure 6), participants indicated a subjective preference for captions without any confidence information.

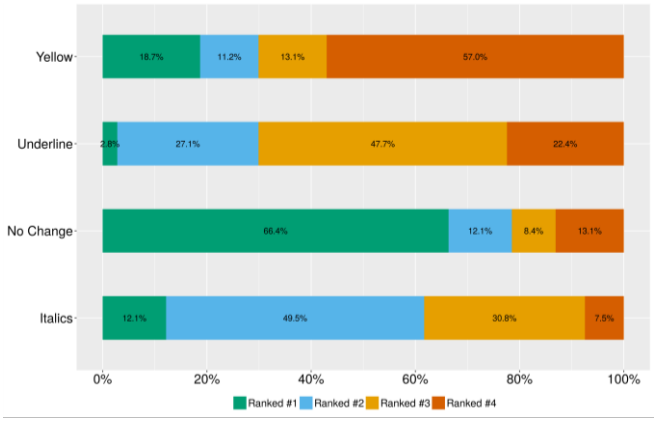



**Figure 9. Larger Study: Ranking of Markup Types**

The findings in this section primary address research question RQ3 from Section 2: "When viewing captions of speech in a one-on-one meeting, which confidence markup do DHH users subjectively prefer?" In addition, the results presented for the question about distraction partially address research question RQ2. Overall, participants expressed a preference for captions without any such markup. In Section 5, we speculate why this might be the case.

# 4. QUALITATIVE RESULTS

This section describes how we analyzed the open-ended responses collected during the larger study using a qualitative methodology. After collecting text comments from 107 participants during the larger study, we obtained a total of 364 individual comments (a total of 6,112 words) from those whom responded – after we had excluded responses such as "no comment" or equivalent. Following the methodology of [40], we began with a round of open-coding in which two annotators independently coded the texts. After meeting to discuss their independent codes, a revised code list was produced, and the annotators formed a consensus coding. One annotator was a Deaf native ASL signer, and the other was a hearing individual: this was done intentionally to reflect the multiple perspectives that are relevant during a meeting between a DHH user and a hearing user. Next, we performed affinity diagramming and thematic analysis of the data to identify common themes for our research questions. The annotators separately considered what themes were relevant to our research questions before meeting again to form the final list. Below, major findings from this analysis are presented, grouped in subsections that correspond to our four research questions from Section 2.

## 4.1 RQ1: Initial view of ASR captioning

At the beginning of the study, before the participants watched the videos, we asked them if they had prior experience with ASR, as we wanted to know if they were already using ASR in their daily lives. Approximately 20% of our participants had prior experience with ASR tools, which may indicate some interest (but significant

opportunity for growth) among DHH users. For those that answered affirmatively, we then followed up with a question, "If you selected 'Yes' please explain your experience with ASR." Our coding of this data indicated that participants had both **positive experiences** and **negative experiences** with ASR. Several who reported a positive experience indicated that they appreciated the fact that they could communicate with hearing people or understand spoken content:

> "I did use my iphone app to pick up news like CNN since sometime CNN do not provide closed caption on live tv" (P97)
>
> "AVA app and it was great! most of the time accurate and it becomes funny when it has errors." (P27)

Among those that shared negative experiences, a common theme that arose was that they were **dissatisfied with ASR accuracy** and they felt that it was too **frustrating** to follow the captioning. Participants believed that they **noticed when there were errors**:

> "I believe they have this on youtube and such it becomes really annoying when the captions are all messed up because i really want to understand what is being said" (P88)
>
> "The hearing participate communicated through ASR app on the phone with me. It was quite interesting even there were several errors and possibly barriers in the communication between us." (P79)

## 4.2 RQ2: Key Issues in Assistive ASR Tools

Our second research question centered on what issues or factors DHH users believed to be important when using ASR captioning as an assistive tool. Our survey included the question: "Could you explain your perspective on computer captioning as an assistive technology?" As mentioned in the previous research question and discussed further in Section 4.3.2, the most frequent issue (N=43) that participants discussed was **ASR accuracy** and how it directly impacted their ability to understand spoken content. Other than accuracy, participants mentioned that they were resistant to using ASR for a variety of reasons, including: concern about **replacing ASL interpreting**, the **lack of bidirectional communication**, and lacking **reliable access** to technology for impromptu ASR interactions. For example:

> "It can be beneficial. I fear that people would abuse the use of assistive technology as it will not be fully reliable. Using a live interpreter will always have my vote." (P14)
>
> "I do support the idea of ASR but there is one issue.. it help us to understand what hearing people are saying but how can we input our message to them through ASR. I cant even speak at all and how can they receive information from me?" (P96)
>
> "I think it is a wonderful technology but the situation is that sometimes people doesn't carry their computers if something comes up." (P81)

## 4.3 RQ3: Preference of Confidence Display

For each of the four markup styles presented in the larger study, we asked participants to discuss their opinion via an open-ended question: "Do you have any comments about this captioning style?" In the following subsections, we will discuss major themes that emerged from the participants' responses to this question (ordered by frequency, beginning with the most frequent issue below).

### *4.3.1 Distraction from Markup*

By far, the most common issue (N=32) raised by participants was that they felt like they could not focus on the information content of the "business meeting" scenario video because the markup style was **too distracting**. It is important for the reader to note that the numerical information in the output of ASR that indicates its confidence on individual words can be wrong: incorrect words for which the ASR was "confident, yet wrong" or correct words for which the ASR was "uncertain, but actually correct." When reading captions with markup, the participants had to cognitively process the markup, the word, and the overall sentence context before they could figure out whether they should trust the captioning (or if they needed to guess what was actually spoken). Many participants indicated that this activity was mentally taxing and unpleasant:

> "In this case this change in fonts is distracting because sometimes styled words are correct sometimes they are not. So I do not know which is correct or not. I also notice some mistakes in non-styled words. In this case the change in styles does not serve its purpose." (P11, *Italics*)
>
> "It is actually distracting if it highlights so many words that are pointless to be acknowledged about what speaker was talking about." (P43, *Underline*)
>
> "Honestly I don't like the highlighted words because it's very distracting and it can be misleading to read. So I wouldn't recommend this kind of captioning style." (P68, *Yellow*)

### *4.3.2 Perception of Accuracy when Errors Visible*

The responses indicated that the majority of participants realized that when markup was added to the captioning, they could readily spot errors. However, it appeared that increasing the visibility of the errors led participants, in some cases, to believe that the captions with markup were less accurate (relative to captions without markup). Although there was no actual difference in the ASR accuracy rate across the different markup conditions, the use of visible markup seemed to increase the ***perceived inaccuracy***, which could reduce acceptance of this technology among users (N=21).

> "The caption was good but some mistake words and no period or comma" (P100, *Italics*)
>
> "however the video in this research did not underlined incorrect words. Not sure if it is a system error or what" (P45, *Underline*)
>
> "It does point out when something is typed wrong and its not what the person said" (P72, *Yellow*)

### *4.3.3 Resistance to Unfamiliar Captioning Markup*

Some (N=13) participants mentioned that they were accustomed to watching movies or television with closed captioning (or subtitles) and wanted consistency of experience. They did not see sufficient benefits of confidence markup to offset this **unfamiliarity**:

> "today we have close caption on the television or movie and their closed caption perfectly smooth" (P6, *Italics*)
>
> "The underline don't need it and I suggest to you used from Netflix it is way to better used CC" (P87, *Underline*)
>
> "White/black only. It is very common closed captions." (P97, *Yellow*)

### *4.3.4 Confusion Regarding the Intent of Markup*

A few (N=7) participants expressed confusion regarding what the markup was supposed to convey. The study had begun with a brief video, in ASL (and a handout containing the script in English), explaining the purpose of the study and what the markup was meant to indicate. Despite this, several participants were **confused about their purpose**, revealing that training may be necessary for use of confidence markup captions:

> "It makes no sense for some words in the captions to be italicized." (P60, *Italics*)

> "I actually got confused with the captions because I was not sure if the underline makes words look important or not." (P64, *Underline*)

> "I think Yellow means it fixes the errors if that is true then I 100% strongly agree!" (P94, *Yellow*)

### *4.3.5 Overall experience of confidence markup*

At the end of the larger study, we asked two questions to elicit opposing perspectives on captioning with confidence markup. The first question was, "Could you list what you liked about the confidence display that you saw today?" The responses had an overall positive tone, even though 37% of participants mentioned that this has been their first time seeing captions with confidence information. The most frequently mentioned themes were: markup leading to more **awareness about errors** in captions, markup helping users understand **how ASR works**, and markups leading to **increased user confidence in ASR as a tool**. For example:

> "It was interesting to see what the captioning wasn't confident about. I didn't know that's what was happening." (P71)

> "bad captions are better than no captions" (P76)

> "I like knowing which one is wrong and which one is right" (P81)

> "It will display accurate word to word. I was amazed about that." (P59)

We then asked a question to elicit more critical feedback, "Could you list some things that should be improved?" As previously discussed in Section 4.3.2, most participants mentioned that if ASR accuracy was improved, they would be more enthusiastic to use the technology. Other than **accuracy**, participants brought up a wide range of issues such as: **styling / appearance** of the captioning, **emphasis on certain words**, and the **learning curve**. Our participants were not able to control the appearance of the captions in this study (e.g. number of lines, font size), yet it is feasible for those issues to be addressed in an actual application. Some participants had unique ideas which could be explored further:

> "I wish the caption could be filled with highlighter so that we cannot miss any important information." (P106)

> "Example of sounds. Showing who is voices. Showing specific of sounds."(P73)

One challenge for the participants was to become acquainted with the concept of reading **imperfect captioning generated by ASR**, as opposed to generated by human (which typically has no errors). This learning curve might not be easy for some DHH users to acquire, but one participant summed it up nicely:

> "Beginning was crazy and awkward but slowly understanding more better." (P94)

## 4.4 RQ4: Other applications of ASR tools

To capture how our DHH participants may imagine themselves using ASR tools they experienced during our study in other contexts, we asked, "Could you list other ways that Automatic Speech Recognition could be used?" Many participants saw potential utility from ASR technology and proposed locations or situations where they thought ASR would be useful: **public places**, **appointments**, **family events**, **travelling**, and **cultural events** (e.g. concerts). Participants indicated a desire to have more access to spoken information, e.g. public announcements. This comment was typical of what we observed from many responses:

> "At an airport when announcing arrived or departed flights automatic speech recognition could come in handy for the deaf and hard of hearing. The same goes for the train bus cruise or any type of station." (P58)

Furthermore, some participants wanted to use ASR as an archiving tool so they could peruse content at their own time and pace:

> "Save the video and can read replay captioning." (P103)

# 5. CONCLUSION

This study has investigated whether it is possible to utilize the confidence values from an ASR engine, to provide DHH users with additional information in an automatic captioning system, for supporting meetings between DHH and hearing users. Through a pair of studies with 128 total DHH participants, we have compared multiple forms of visual display to present this confidence information, and we collected subjective preferences from DHH users on each. In addition, after asking DHH participants to engage in a simulation of a one-on-one meeting with a hearing person, with automatically generated captions that included confidence markup, we surveyed participants about their opinions on this technology. Quantitative and qualitative analysis of these responses suggested several major findings, as discussed in Sections 3 and 4 above:

- Participants **expressed interest** in adding markup to captioning to indicate the words for which ASR was uncertain.
- However, after experiencing such captions during a simulated meeting, participants expressed subjective preferences for captions that **did not include** any confidence markup.

We speculate various potential explanations for this contrast:

- Participants may have disliked the specific set of visual markup display approaches used in the larger study.
- Participants did not consciously notice any benefit from the additional confidence information, e.g. perceiving that any benefits were outweighed by the potential for distraction.
- Participants' prior experience with captions without confidence markup led to their subjective preference for a familiar technology, and the exposure to other conditions during this simulation was too brief for them to learn how to utilize this information. It is conceivable that if standard methods of confidence display were widely adopted, then users may gain familiarity in interpreting and utilizing this information in ASR captions.
- The simulated nature of the one-on-one meeting activity in the larger study did not adequately capture real-word aspects of an interactive meeting with a hearing conversational partner, thereby obfuscating potential benefits of this technology.
- Potential benefits of this technology may only occur under specific conditions, e.g. particular ASR accuracy rates, text reading complexity, or participant literacy levels – as discussed in Section 6 below, a wide variety of settings for these parameters would need to be examined in future work.

Our qualitative analysis of the feedback from participants also led to several findings about factors that DHH consider important in automatic captioning and useful applications that users foresee:

- **Accuracy** was salient for DHH users considering the potential of ASR captioning; additional concerns included the potential **loss of ASL interpreting services**, supporting **bidirectional** communication, and the **reliability** of the technology.
- In regard to the confidence markup they experienced in the study, participants were concerned about **distraction** caused from changes in text appearance. Some commented on how markup made **errors more apparent** and how changes in text appearance during captions were **unfamiliar**. A few

participants **did not understand the purpose** of markup, which may indicate a need for more introduction or training.
- Participants suggested potential **applications** for ASR-generated captions, e.g. for announcements in public places, appointments, family gatherings, or cultural events.

The overall aim of this study has been to explore the potential of ASR captioning technology for use in meetings, especially whether conveying confidence may benefit DHH users. The goal of our analysis has been to distill our findings into a form that may be useful for future researchers interested in this application. In particular, by identifying key concerns of DHH participants, it may be possible to explore alternative designs for this technology to mitigate perceived deficiencies or address key concerns.

# 6. LIMITATIONS AND FUTURE WORK

Section 1.3 described the Word Error Rate (WER) of the ASR-generated captions in our study, but ASR accuracy has been improving in recent years, e.g. [19, 46]. This study did not examine how variation in WER might have influenced participants' preference scores of confidence markup or influenced their views on the use of ASR for captioning. The findings of this study may be specific to particular error rates in caption text, and a future study would be needed to examine this issue.

In future work, it may be useful to consider the intersection of literacy and text complexity. Most of our participants were students on a university campus, and it is known that most deaf college students have reading levels below that of sixth grade (11-12 years old) [1, 33]. To analyze the reading complexity of our stimuli, we analyzed them using the Flesch-Kincaid [14] formula: the Flesch Reading-Ease Score was (FRES: mean=69.76, median=69.21 with SD=9.69) and the Flesch-Kincaid Grade Level was (FKGL: mean=8.6, median=8.73 with SD=2.28). Based on those scores, the text appearing in the captions of our video stimuli was at the 8$^{th}$ grade reading level (13-14 years old) approximately. In future work, we may examine stimuli with a range of text reading complexity (and measure the literacy rate of our participants on a standardized test) to investigate if these factors may influence participants' preferences in an ASR based captioning system.

While this study employed subjective preference questions and open-ended feedback to probe DHH participants' views of ASR captioning technology, additional measurements are possible:

- Prior work has utilized eye-tracking technology to study how DHH participants utilize captions [41] and how modifications to text appearance influences their eye movements [37]; we may employ eye-tracking on ASR captioning for meetings and the use of confidence markup in future work.
- Our larger study included comprehension questions to encourage participants to engage in the simulated meeting task. While none of our research questions in this study focused on potential changes in users' comprehension scores based on different markup types, out of curiosity, we conducted an ANOVA to compare comprehension scores for different markup types, but we found no statistically significant differences, similar to [39]. This may illustrate the difficulty in measuring comprehension differences from displaying captions with different forms of confidence markup. Interestingly, while participants indicated that some markup styles were distracting (Section 3), there were no differences in comprehension scores in this study. We plan to investigate comprehension scores in a future study, considering the methodology of [4], who measured DHH students' comprehension of an educational lecture.

# 7. ACKNOWLEDGEMENTS

This work has been supported by a Google Faculty Research Award and by the National Technical Institute for the Deaf (NTID). Kasmira Patel and Anmolvir Kaur assisted with data collection.

# 8. REFERENCES

[1] J. Albertini, C. Mayer. 2011. Using miscue analysis to assess comprehension in deaf college readers. *The Journal of Deaf Studies and Deaf Education*, 16(1):35.

[2] Consumer Electronics Association. 2013. Cea-708-e (ansi): Digital television (dtv) closed captioning.

[3] F. G. Bowe. 2002. Deaf and hard of hearing Americans' instant messaging and e-mail use: A national survey. *American Annals of the Deaf*, 147(4):6–10.

[4] A. Brandão, H. Nicolau, S. Tadas, V.L. Hanson. 2016. Slidepacer: A presentation delivery tool for instructors of deaf and hard of hearing students. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '16), ACM, New York, NY, USA, 25–32. DOI: http://dx.doi.org/10.1145/2982142.2982177

[5] CaptionMax. 1993. *Suggested Styles and Conventions for Closed Captioning*. WGBH Educational Foundation, Boston.

[6] Federal Communications Commission. 2015. *Video relay services.* http://www.fcc.gov/guides/video-relay-services

[7] M. Crabb, R. Jones, M. Armstrong, C.J. Hughes. 2015. Online news videos: The UX of subtitle position. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (ASSETS '15), ACM, New York, NY, USA, 215-222. DOI: http://dx.doi.org/10.1145/2700648.2809866

[8] M. Davoudi, M. B. Menhaj, N. S. Naraghi, A. Aref, M. Davoodi, M. Davoudi. 2012. A fuzzy logic-based video subtitle and caption coloring system. *Advances in Fuzzy Systems*. DOI: http://dx.doi.org/10.1155/2012/671851

[9] DCMP. 2016. DCMP Captioning Key - Quality Captioning. http://www.captioningkey.org/quality_captioning.html

[10] L. Elliot, M. Stinson, J. Mallory, D. Easton, M. 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* (ASSETS '16), ACM, New York, NY, USA, 271-272. DOI: http://dx.doi.org/10.1145/2982142.2982198

[11] L.B. Elliot, M.S. Stinson, B.G. McKee, V. S. Everhart, P.J. Francis. 2001. College students' perceptions of the c-print speech-to-text transcription system. *Journal of deaf studies and deaf education*, 6(4):285–298.

[12] V. Ferdiansyah, S. Nakagawa. 2013. Effect of captioning lecture videos for learning in foreign language. In *Proc. SLP Meeting of Info. Processing Society of Japan*. SLP-97, No. 3.

[13] D.R. Flatla, C. Gutwin. 2012. Situation-specific models of color differentiation. *ACM Trans. Access. Comput.* 4, 3, Article 13, 44 pages. DOI: http://dx.doi.org/10.1145/2399193.2399197

[14] R. Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221-223.

[15] Y. Gaur, W.S. Lasecki, F. Metze, J.P. Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th Web for All Conference* (W4A '16). ACM, New York, NY, USA, Article 23, 8 pages. DOI: https://doi.org/10.1145/2899475.2899478

[16] S.R. Gulliver, G. Ghinea. 2003. How level and type of deafness affect user perception of multimedia video clips. *Universal Access in the Information Society*, 2(4):374–386.

[17] R.P. Harrington, G.C. Vanderheiden. 2013. Crowd caption correction (ccc). In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '13). ACM, New York, NY, USA, Article 45, 2 pages. DOI: http://dx.doi.org/10.1145/2513383.2513413

[18] R. Hong, M. Wang, M. Xu, S. Yan, T.-S. Chua. 2010. Dynamic captioning: video accessibility enhancement for hearing impairment. In *Proceedings of the 18th ACM international conference on Multimedia* (MM '10), ACM, New York, NY, USA, 421–430. DOI: http://dx.doi.org/10.1145/1873951.1874013

[19] X. Huang, J. Baker, R. Reddy. 2014. A historical perspective of speech recognition. *Comm. ACM*, 57(1):94–103.

[20] D.W. Jackson, P.V. Paul, J.C. Smith. 1997. Prior knowledge and reading comprehension ability of deaf adolescents. J*ournal of Deaf Studies and Deaf Education*, 2(3):172–184.

[21] J. Jankowski, K. Samp, I. Irzynska, M. Jozwowicz, S. Decker. 2010. Integrating text with video and 3d graphics: The effects of text drawing styles on text readability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10). ACM, New York, NY, USA, 1321-1330. DOI: http://dx.doi.org/10.1145/1753326.1753524

[22] L. Jefferson, R. Harvey. 2006. Accommodating color blind computer users. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility* (Assets '06). ACM, New York, NY, USA, 40-47, DOI: http://dx.doi.org/10.1145/1168987.1168996

[23] S. Kawas, G. Karalis, T. Wen, R.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* (ASSETS '16). ACM, New York, NY, USA, 15-23, DOI: http://dx.doi.org/10.1145/2982142.2982164

[24] R. Kheir, T. Way. 2007. Inclusion of deaf students in computer science classes using real-time speech transcription. In *ITiCSE'07*, ACM, NY, NY, USA, 261-265, DOI: http://dx.doi.org/10.1145/1268784.1268860

[25] R.S. Kushalnagar, G.W. Behm, A.W. Kelstone, S. Ali. 2015. Tracked speech-to-text display: Enhancing accessibility and readability of real-time speech-to-text. In *Proc ASSETS'15*, ACM, NY, NY, USA, 223-230, DOI: http://dx.doi.org/10.1145/2700648.2809843

[26] R.S. Kushalnagar, W.S. Lasecki, J.P. Bigham. 2014. Accessibility evaluation of classroom captions. *ACM Trans. Access. Comput.* 5, 3, Article 7, 24 pages. DOI: http://dx.doi.org/10.1145/2543578

[27] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, P. Wolf. 2003. The CMU sphinx-4 speech recognition system. In *Proc ICASSP'03,* 2–5.

[28] D.G. Lee, D.I. Fels, J.P. Udo. 2007. Emotive captioning. *Computers in Entertainment (CIE)*, 5(2):11.

[29] M. Marschark, G. Leigh, P. Sapere, D. Burnham, C. Convertino, M. Stinson, H. Knoors, M. P. Vervloed, W. Noble. 2006. Benefits of sign language interpreting and text alternatives for deaf students' classroom learning. *Journal of deaf studies and deaf education*, 11(4):421–437.

[30] D. Miller, K. Gyllstrom, D. Stotts, J. Culp. 2007. Semi-transparent video interfaces to assist deaf persons in meetings. In *Proc ACM-SE 45,* ACM, NY, NY, USA, 501-506. DOI: http://dx.doi.org/10.1145/1233341.1233431

[31] M. R. Mirzaei, S. Ghorshi, M. Mortazavi. 2012. Using augmented reality and automatic speech recognition techniques to help deaf and hard of hearing people. In *Proc VRIC'12*, ACM, NY, NY, USA, Article 5, 4 pages. DOI: http://dx.doi.org/10.1145/2331714.2331720

[32] H. Nambo, S. Seto, H. Arai, K. Sugimori, Y. Shimomura, H. Kawabe. 2012. Visualization of non-verbal expressions in voice for hearing impaired: Ambient font and onomatopoeic subsystem. In *Proc ICCHP'12 volume 1*, 492–499.

[33] S.J. Parault, H.M. Williams. 2010. Reading motivation, reading amount, and text comprehension in deaf and hearing adults. *J. of Deaf Studies and Deaf Educ.*, 15(2):120-135.

[34] A. Piquard-Kipffer, O. Mella, J. Miranda, D. Jouvet, L. Orosanu. 2015. Qualitative investigation of the display of speech recognition results for communication with deaf people. In *SLPAT'15,* 36-41. Assoc. of Comp. Linguistics.

[35] S.S. Prietch, N.S. de Souza, L.V.L. Filgueiras. 2014. A speech-to-text system's acceptance evaluation: would deaf individuals adopt this technology in their lives? In *UAHCI'14,* 440–449.

[36] R. Rashid, Q. Vy, R.G. Hunt, D.I. Fels. 2007. Dancing with words. In *Proc C&C'07*, ACM, NY, NY, USA, 269–270. DOI: http://dx.doi.org/10.1145/1254960.1255007

[37] K. Rathbun, L. Berke, C. Caulfield, M. Stinson, M. Huenerfauth. 2017. Eye movements of deaf and hard of hearing viewers of automatic captions. Journal on Technology & Persons with Disabilities, 5, 130-140, http://hdl.handle.net/10211.3/190208

[38] A. Secarã. 2011. RU ready 4 new subtitles? Investigating the potential of social translation practices and creative spellings. In M. O'Hagan (ed.), *Translation as social activity: Community translation 2.0 [Special issue]. Linguistica Antverpiensia New Series*, 10, 153–172.

[39] B. Shiver, R. Wolfe. 2015. Evaluating alternatives for better deaf accessibility to selected web-based multimedia. In *Proc ASSETS'15*, ACM, New York, NY, USA, 223–230. DOI: http://dx.doi.org/10.1145/2700648.2809843

[40] Strauss, A., & Corbin, J. (1998). Basics of qualitative research: Techniques and procedures for developing grounded theory (2nd ed.). Thousand Oaks, CA: Sage.

[41] A. Szarkowska, I. Krejtz, Z. Klyszejko, A. 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):363–378.

[42] K. Vertanen, P.O. Kristensson. 2008. On the benefits of confidence visualization in speech recognition. In *Proc CHI'08*, ACM, New York, NY, USA, 1497–1500. DOI: http://dx.doi.org/10.1145/1357054.1357288

[43] Q.V. Vy. 2012. Enhanced captioning: speaker identification using graphical and text-based identifiers. *Theses and dissertations.* Paper 1702.

[44] M. Wald. 2005. Using automatic speech recognition to enhance education for all students: Turning a vision into reality. In *Proc IEEE FIE'05*, S3G-S3G. IEEE.

[45] F. Wang, H. Nagano, K. Kashino, T. Igarashi. 2015. Visualizing video sounds with sound word animation. In *Proc ICME'15*, 1–6. IEEE.

[46] W. Xiong, J. Droppo, X. Huang, F. Seide, . Seltzer, A. Stolcke, D. Yu, G. Zweig. 2016. Achieving human parity in conversational speech recognition. *Computing Research Repository (CoRR),* http://arxiv.org/abs/1610.05256

[47] G.N. Yannakakis, J. Hallam. 2011. *Ranking vs. Preference: A Comparative Study of Self-reporting*, In: S. D'Mello, A. Graesser, B. Schuller, J.C. Martin JC (eds) *Affective Computing and Intelligent Interaction. ACII 2011. Lecture Notes in Computer Science,* 6974. Springer, Berlin.