Design and Evaluation of Accessible Collaborative Writing Techniques for People with Vision Impairments

MAITRAYE DAS, Northwestern University
ANNE MARIE PIPER, University of California, Irvine
DARREN GERGLE, Northwestern University

Collaborative writing tools have been used widely in professional and academic organizations for many years. Yet, there has not been much work to improve screen reader access in mainstream collaborative writing tools. This severely affects the way people with vision impairments collaborate in ability-diverse teams. As a step towards addressing this issue, the present paper aims to improve screen reader representation of collaborative features such as comments and track changes (i.e., suggested edits). Building on our formative interviews with 20 academics and professionals with vision impairments, we developed auditory representations that indicate comments and edits using non-speech audio (e.g., earcons, tone overlay), multiple text-to-speech voices, and contextual presentation techniques. We then performed a systematic evaluation study with 48 screen reader users that indicated that non-speech audio, changing voices, and contextual presentation can potentially improve writers' collaboration awareness. We discuss implications of these results for the design of accessible collaborative systems.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; **Empirical studies in accessibility**.

Additional Key Words and Phrases: Collaborative writing, vision impairments, non-speech audio, screen readers, ability-diverse collaboration

ACM Reference Format:
Maitraye Das, Anne Marie Piper, and Darren Gergle. 2021. Design and Evaluation of Accessible Collaborative Writing Techniques for People with Vision Impairments. *ACM Trans. Comput.-Hum. Interact.* 1, 1, Article 1 (January 2021), 42 pages. https://doi.org/10.1145/3480169

1 INTRODUCTION

Collaborative writing has become an integral part of professional and academic work, as business, education, engineering, law, and other organizational sectors are increasingly promoting group work that involves writing reports, papers, and articles together with others [89, 93]. Decades of research in HCI and CSCW has focused on understanding collaborative writing practices [13, 15, 46, 61, 63, 89, 94] and developing theoretical frameworks [39, 49, 66, 77] and experimental systems (e.g., [8, 33, 62]) to meet the needs of collaboration and coordination within teams. In parallel, a multitude of tools (e.g., Google Docs, Microsoft Office 365, Overleaf, etc.) have been designed that brought to fruition ideas from early research in the form of collaborative features such as comments, track changes, revision history, and real-time edit notifications. Researchers have also developed

Authors' addresses: Maitraye Das, maitraye@u.northwestern.edu, Northwestern University, Evanston, IL, USA; Anne Marie Piper, ampiper@uci.edu, University of California, Irvine, Irvine, CA, USA; Darren Gergle, dgergle@northwestern.edu, Northwestern University, Evanston, IL, USA.

1073-0516/2021/1-ART1 $15.00
https://doi.org/10.1145/3480169

ways to visualize how co-authors use and interact with these new collaborative features and how their individual actions and contributions shape the production of a shared document over time [43, 45, 64, 82, 88, 89, 97].

Despite significant academic and commercial interest in collaborative writing systems, less is known about how these systems support teams of people with diverse physical, cognitive, or sensory disabilities. Our focus is specifically on ability-diverse teams that involve people with vision impairments working with sighted colleagues and the ways in which the design of collaborative tools and features can support collaborative writing activities that are distributed across time and space. Our work identifies key design challenges associated with using screen readers to perform collaborative writing and systematically evaluates new auditory representations of collaborative features to address these challenges. We focus on accessibility issues in asynchronous collaborative writing where co-authors work on a shared document one at a time. While recent developments in collaborative writing tools offer many opportunities for synchronous collaboration (i.e., multiple authors working on a document simultaneously in real-time), many people still use asynchronous editing features, such as comments and suggested edits, to write together and exchange feedback [13, 89]. As such, improving screen reader access for asynchronous editing is an important first step towards ensuring accessible collaboration in ability-diverse teams.

We ground the design and study of novel auditory representations of collaborative writing features in interviews with 20 academics and professionals with vision impairments who regularly perform collaborative writing using screen readers. Our prior work reports findings from these interviews that highlight the ways in which visually impaired writers navigate through an ecosystem of tools comprised of multiple word processors and screen readers, negotiate accessibility needs with sighted collaborators, and face broader social, professional, and organizational challenges in ability-diverse collaboration [29]. In the current paper, we report new aspects of the interview data that detail the complexities visually impaired writers encounter when using collaborative features (e.g., comments and edits) during asynchronous collaborative writing. Our current analysis of interview data reveals that screen reader users face four key challenges as part of developing and maintaining collaboration awareness [30] (i.e., understanding who did what and where) in a shared document: (1) distinguishing between document content, collaboration markup, and comments/edits from others, (2) understanding how document content evolves through underlying edits, (3) managing disruption in workflow created by verbose spoken announcements for collaboration markup, and (4) controlling the influx of collaboration information.

To address these challenges identified through our interview study, we designed and developed a variety of auditory representations that incorporate non-speech audio (e.g., earcons [35] and tone overlays), multiple text-to-speech voices, and contextual presentation techniques. The auditory representations were designed to help writers identify three key pieces of information that facilitate collaboration awareness in asynchronous editing: *where the comments are*, *who commented what*, and *who edited what*. We evaluated these techniques through a within-subjects experiment with 48 visually impaired writers who frequently perform collaborative writing activities using screen readers. Our results indicated that non-speech audio, changing voices, and contextual presentation techniques are promising approaches for improving collaboration awareness among screen reader users. We found that tone overlay works as the least disruptive approach to understanding where comments are located while simultaneously comprehending the text content, specifically in complex passages with densely populated and overlapping comments. Similarly, reading collaborators' edits or comments in different voices makes it easier to keep track of who edited or commented about a specific text segment and what they said in their comments, although this benefit diminishes when more collaborators contribute to a document. Additionally, presenting edits in the context of a sentence helps people in figuring out how the sentence evolved after multiple edits.

This paper makes three key contributions to the fields of HCI, CSCW, and accessible computing. First, we contribute deeper empirical understandings of the complexities of how screen reader users maintain collaboration awareness during asynchronous writing, which extends prior work on how blind and sighted people collaborate in professional [21, 87], educational [53, 55, 56, 76, 79], creative work [16, 28, 71], and everyday living contexts [20, 91, 95]. Second, our systematic evaluation contributes new insights regarding how screen readers and word processors can better support collaborative writing through contextual markers and non-speech audio cues – techniques that have previously been used to improve non-visual access to graphical interfaces [52, 60, 74], diagrams [51, 54, 80], and navigation [34, 92] for people with vision impairments. Third, we synthesize our findings from across the two studies to highlight design tradeoffs and considerations for enhancing accessibility in future collaborative writing systems used by blind and sighted teams.

2 RELATED WORK

Our work is informed by research on collaborative writing tools and practices, accessibility of writing tools and dynamic interfaces as well as the use of non-speech audio representations in assistive technology.

2.1 Collaborative Writing Tools and Practices

Over the years, HCI and CSCW scholars have investigated how to design collaborative writing systems to support co-authors and how people produce shared documents, exchange feedback, and interact with each other using these tools [13, 15, 46, 61, 63, 94]. Researchers have developed experimental systems (e.g., ShrEdit [62], Quilt [33], and SASSE [8]), theoretical frameworks [39, 49, 66, 77], and have accumulated empirical knowledge of collaborative writing practices that led to further improvement of widely available collaborative systems [13, 30, 89, 94]. For example, Dourish and Bellotti put forth the concept of *collaboration awareness*, or the work of understanding who did what, where, and when to coordinate group efforts within a shared document [30]. In addition to developing collaboration awareness, Birnholtz and Ibara found that people also paid attention to how other co-authors might interpret their actions in the document and, subsequently, left comments explaining those actions [13]. To better support such group dynamics, these researchers suggested that collaborative writing tools incorporate a "suggestion mode" where one's edits on others' text are shown as suggested edits [14] – a feature that was later implemented on Google Docs. In a separate study, Wang et al. illustrated that while performing synchronous collaborative writing using Google Docs, users manually highlighted text written by each author using unique colors or fonts to develop retrospective awareness of authorship attribution, which was not readily available through real-time editing cursors or revision history features [89]. Collectively, this research focuses on understanding what kinds of collaboration information people need to learn and convey to others as they write together and how to design features to meet those needs.

In a similar vein, researchers have also developed summary visualization techniques to demonstrate who contributes what in a shared document and how the document evolves over time. Wang et al. aggregated streams of revision history data on Google Docs and developed two systems – DocuViz [88], which visualizes collaboration patterns in a group, and AuthorViz [89], which color-codes each author's edited text in the final version after dozens of revisions (both systems were later implemented as Google Chrome extensions). Zhu et al. designed CEPT, a collaborative editing platform that facilitates language knowledge sharing among non-native speakers by presenting aggregated edits of multiple co-authors and allowing users to incorporate others' edits into their writing [97]. A separate thread of work has also explored ways to represent the intricacies of various co-authors' actions that reveal the complex interdependencies and coordination performed over a shared document. For instance, researchers have developed dynamic and interactive

visualizations to show the organization and hierarchical structure of the collaborative text [64], quality of co-authors' contributions [82], location of their gaze within a real-time editor [43], and temporal, spatial, and territorial nature of their revision patterns [45, 83]. Overall, this body of work illustrates how visualization techniques can support people in effectively and efficiently consuming collaboration information on individual and aggregate levels. In this paper, we focus on a relevant but distinct problem: we design and evaluate *auditory representations* of collaboration information to support visually impaired writers who use screen readers to access collaboration information.

2.2 Accessibility of Writing Tools and Dynamic Web content

While much research has been conducted on collaborative writing practices, applications, and visualizing collaboration information, less attention has been given to supporting collaborative writing for teams involving people with vision impairments. Prior work that focused on exploring accessibility issues in collaborative writing highlighted that screen reader users find it extremely challenging to navigate, perceive, and interact with collaborative features (e.g., comments, track changes, real-time editing) that are available on common writing platforms such as Microsoft Word and Google Docs [26, 29, 73]. Even basic functionality, such as formatting and resizing documents, understanding table content, searching text, and traversing menu options were also difficult to access via screen readers at the time of this research [26, 27, 57]. To address this, researchers have developed Microsoft Word or Google Docs extensions to improve accessibility of basic document editing and formatting features [26, 57] and provide additional context for some collaborative features such as track changes [73]; yet, the design of these techniques has not been explored in depth or evaluated systematically.

Beyond accessibility of these writing tools, researchers have also studied the challenges people with vision impairments face in accessing dynamic web content and how they cope with these challenges [11, 12, 17, 50, 85]. While dynamic web interfaces and collaborative writing systems present two different contexts of use, there are similarities in the way information is represented and updated dynamically on these interfaces (e.g., real-time editing notifications appear as co-authors write together). As such, screen reader users often face similar challenges in accessing dynamic web content as they do in collaborative writing tools. For example, often users cannot find their desired information amidst dynamically updated web content that may or may not be relevant to them [17]. In many cases, they may not even be able to clearly identify whether their desired content is inaccessible or does not exist at all [11]. To address these issues, researchers have developed various systems (e.g., [18, 24, 72]) and guidelines for best practices [3, 50, 68] over the years. For instance, Brown et al. proposed tailored presentation of web updates, where the auditory browser only triggers non-speech sound alerts for automated updates but provides verbal descriptions for user-initiated updates that make meaningful changes in the webpage [24]. Sato et al. augmented the sequential model of web navigation by supplementing the primary voice output with a secondary whisper of contextually relevant information [72]. Recently, researchers have started investigating faster skimming of web content non-visually and proposed a number of techniques that include hierarchical views with dynamically generated outlines of web content [96], personalized adaptations according to individual preferences inferred from browsing history [6, 96], and automated browsing actions through semantic web modeling [7]. In our work, we expand on this literature by focusing on improving accessibility and usability of collaborative writing systems. Specifically, we study whether non-speech audio and contextual presentations that have been found useful in presenting dynamic web content [24, 72] can also enhance the way screen reader users extract and consume collaboration information in a shared document.

2.3 Non-speech Audio Representations in Assistive Technology

There is an extensive literature on non-speech auditory representations that convert graphical, textual, and visual data to representative sound (e.g., auditory icons, musicons, and auditory emoticons) or structured and abstract sound (e.g., earcons, spearcons, spatial sound, and sonification) (for an overview, see [4, 35]). One of the most important benefits of non-speech audio is that it can leverage users' auditory perception skills to communicate information in unique ways, whereas explicit speech representations can be intrusive, time consuming, and socially unacceptable in many situations [35]. In particular, for people with vision impairments, non-speech audio representations have become a key technique in designing accessible and assistive systems. Researchers and industry practitioners have been developing auditory interfaces and augmentations to support navigation and wayfinding for visually impaired individuals using spatial binaural sound [2, 34]. Relatedly, Tomlinson et al. found that auditory graphs that alter sound attributes (e.g., pitch) and use non-speech sounds to represent numerical values and graph features can help improve visually impaired students' engagement in the class [80]. Sonification techniques were also helpful for an avid blind gamer in understanding the speed and trajectory of a car, and direction, sharpness, length, and timing of upcoming turns in a racing game [74]. Mendes et al. used spatial sound and multiple text-to-speech voices to support workspace awareness of blind users in collaborative tabletop activities [52]. Similarly, Metatla et al. developed multimodal interfaces that allowed blind and sighted co-workers to explore diagrams through simultaneous visual and audio-haptic representations including earcons, synthesized speech, and magnetic force feedback [54]. Others have combined a number of sonification techniques and non-speech audio with synthesized speech feedback to support people with vision impairments in proximity estimation [34, 92], recognizing shapes [51], learning touchscreen gestures [60], receiving word completion suggestions [59], identifying programming constructs [41], and perceiving video annotations [32] and image descriptions[10].

Perhaps what makes auditory representations so pervasive within assistive technology research is that people with vision impairments often outperform their sighted peers in listening abilities. Researchers found that visually impaired users, specifically, early adopters of screen readers, can comprehend synthetic speech at a much higher rate compared to sighted individuals [19, 75]. Blind users can also identify and understand relevant content when presented through concurrent synthesized voices [38]. Also, they are generally adept at different auditory tasks such as pitch identification [40], sound localization [86], and remembering audio stimuli [69].

Overall, auditory enhancements and synthesized speech effects have long been explored in assistive technology research and products to present complex graphical information quickly and efficiently for people with vision impairments [4, 35, 74]. In this paper, we analyze the challenges of collaboration awareness for screen reader users when writing collaboratively and investigate whether various auditory representations that leverage non-speech audio and contextual presentation techniques effectively support understanding collaborative information.

3 FORMATIVE STUDY: METHOD

We conducted interviews with visually impaired writers to understand their collaborative writing practices and challenges associated with collaborative tools. This study was approved by the Institutional Review Board of Northwestern University. Our prior work [29] reports data from these interviews that address higher level collaboration strategies and group work practices in ability-diverse teams. In this paper, we present additional findings from these interviews regarding how visually impaired writers encounter the lower level features and technical aspects of accessibility in collaborative writing systems.

3.1 Participants

We performed semi-structured interviews with 20 professionals and academics with vision impairments (age ranging between 20-50 years, 9 identified as male and 11 as female). Participants were recruited through our research network and snowball sampling. Most participants live in the United States except David and Grace.[1] All participants perform collaborative writing frequently, except Kaylee who occasionally does so. For screen readers, participants mainly use JAWS, NVDA, and VoiceOver. A few participants also have experience using Microsoft Narrator, Android Talkback, and Google ChromeVox. Participants primarily rely on auditory speech output of screen readers, although a few participants occasionally use braille displays along with screen readers. Interviews specifically focused on how participants perform collaborative writing using auditory speech output of screen readers. Microsoft Word and Google Docs are the most common tools for collaborative writing among the participants. Participants come from different professional backgrounds and most of them perform writing activities with both sighted and blind collaborators. See Table 2 in Appendix A for participant details including their self-reported visual ability, occupation, and the kinds of documents they produce through collaborative writing.

3.2 Procedure

We conducted the interviews remotely through phone or audio/video conferencing tools such as Zoom, Skype, or Facetime as preferred by the participants. All interviews were conducted by the first author between January 2019 and March 2019. At the beginning of each interview, we collected verbal consent from the participants. Interviews lasted for approximately 40-75 minutes. Participants were compensated with US$30 gift cards for their time and effort. All the interviews were audio recorded and transcribed for analysis. We conducted the interviews in a semi-structured format so that participants could freely talk about their experiences interacting with various collaborative writing tools. Our broader interview protocol focused on understanding how visually impaired writers perform group work in predominantly sighted workplaces, with an emphasis on how they communicate their accessibility needs to sighted collaborators, how their collaborators view and act on these accessibility needs, and how group members adapt their work practices to create access. In addition to exploring these broader collaboration practices, we also asked about how visually impaired writers individually interact with collaborative features (e.g., comments and edits) using screen readers and discussed ideas for improving the design of these tools. Our analysis here focuses on responses to the latter interview questions whereas our earlier publication [29] details findings related to higher level collaboration strategies and group work practices.

3.3 Data Analysis

We followed a reflexive thematic analysis method for data analysis [22, 23]. We began by having the first two authors read the interview transcripts. The first author re-read and open coded the interview transcripts with a particular focus on investigating participants' individual interaction with collaborative writing systems. Next, she wrote analytic memos on the codes and collated them into preliminary themes through a process of iteratively comparing data to data and data to emerging themes. All the authors regularly discussed the themes and codes as a group. Finally, we organized the themes according to tasks and goals that participants seek to accomplish while performing collaborative writing, focusing on asynchronous writing activities and the process of collaboration awareness [30]. Each theme captures how participants interact with different collaborative features to achieve specific goals, what complexities arise in this process, how they

[1] All names are pseudonyms.

cope with these complexities, and what design changes could potentially improve their collaborative writing experience.

Although we focus here on the technological features of collaborative systems and how they affect access, our analysis is informed by Kafer's political/relational model of disability [42] and work from other feminist disability scholars [31, 48, 90]. Specifically, we view access as a continuous process that is negotiated through particular socio-material configurations instead of located entirely in the individual, society, or technology. In this work, we focus on improving the design of technological features as a step towards advancing accessibility in group work. However, we recognize that accessibility, and how it is achieved in group interaction, is a complex sociotechnical phenomenon.

4 FORMATIVE STUDY: FINDINGS

Collaborative writing is a complex process that requires co-authors to remain aware of each other's actions, such as who is editing or commenting what, where, and when, and how the shared document is evolving through these actions [13, 30, 39, 77]. Existing collaborative writing applications offer a number of visual cues to help sighted writers remain cognizant of their co-authors' actions in the shared document. For example, in applications like Microsoft Word or Google Docs, comments are juxtaposed in the sidebar beside the document body, text portions where comments are anchored are highlighted, insertions and deletions are represented through underlining and strikethrough, and comments and edits by different co-authors are color-coded. Additionally, various interactive features help sighted people navigate through and respond to collaborators' actions. For instance, when a comment is selected in Word, the saturation of the color highlighting the corresponding text portion increases and the commented text becomes visually prominent; Word also reinforces the connection by converting the dashed line connecting the comment with the anchor text into a solid one (see Figure 1 for an example).

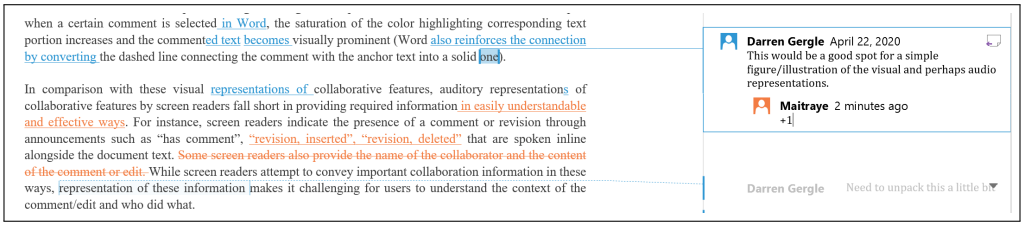

Fig. 1. Screenshot of a Microsoft Word document with edits and comments from two co-authors.

In comparison to these visual representations, screen readers' auditory representations of collaborative features fall short in providing required information in easily understandable and effective ways. For instance, screen readers indicate the presence of a comment or revision by announcing markup phrases such as 'has comment', 'revision, inserted', 'revision, deleted' that are spoken inline alongside the document text. Some screen readers also provide the name of the collaborator and the content of the comment or edit. While screen readers attempt to convey important collaboration information through spoken announcements, the way this information is actually presented makes it challenging for people to perform certain tasks that are essential for developing collaboration awareness. Below we detail these tasks along with the challenges associated with them and potential design changes to alleviate these challenges.

4.1 Distinguishing between Document Content, Collaboration Markup, and Collaborators' Actions

To coordinate group efforts in a shared document, co-authors must learn who edited or commented what and where in the document. As we discussed earlier, screen readers present this collaboration information through serialized spoken announcements where markup phrases (e.g., 'has comment', 'revision inserted') and document content are interlaced. This makes it *"cognitively overloading"* for our participants to differentiate between document content and collaboration markup phrases as well as keep track of different collaborators' actions (e.g., comments/edits). Emma said, *"Track changes tends to muddy the waters very badly. For instance, if I have a document that someone else has changed, I might hear 'the cat deleted rat ate 15 mice changed to'... Some of it is actual text, some of it is deleted text and [I'm] not having a great difference between the different ones."*

Furthermore, screen readers offer no straightforward way to differentiate between hierarchical comments (i.e., replies) and overlapping versus standalone comments. Mike, Isaac, Emma, and Maya explained that they *"try to make sense [of replies and overlapping comments] based on the context of the discussion."* This, however, requires them to go through a *"daunting process"* that involves jumping back and forth between the list of comments and the document text, and performing a number of checking steps. For instance, in the list view of JAWS, each comment appears with a snippet of the text they are attached to. In cases where a comment does not have the snippet of the attached text alongside it, participants *"assume that it's the reply for the previous comment."* Alternatively, if multiple adjacent comments in the sequential list are attached to the same portion of text, those are considered to be overlapping with each other. To avoid this complex procedure, Mike asks his collaborators to reply in a separate comment and to preface their comment with "in reply to your previous comment" instead of leaving replies or overlapping comments.

Participants suggested that one possible way to easily process and distinguish between these intertwined pieces of collaboration information could be using multiple synthesized speech voices or manipulating speech parameters (e.g., pitch). Elena explained, *"There could be more done with sound or pitch or inflection or even using multiple text-to-speech voices to present it... It would be cool if you could set a voice for each of the editors."* Addison even made an analogy between listening to comments/edits from co-authors in different voices and *"see[ing] it in a different color, so you can still keep reading and it doesn't break up your flow."*

While most participants suggested using different voices to elucidate different co-authors' edits/comments, Emma proposed an alternative use-case, where specific voices could denote *"different kinds of text (inserted or deleted text) in a document with track changes."* She emphasized that the characteristics of the voices should semantically align with the type of text they represent. For example, deleted text could be read out in a "deeper" (i.e., low-pitched) voice to delineate that *"that's not really relevant anymore,"* while inserted text could have a *"higher pitched voice so that I could tell, 'Oh, hey, that's new, that wasn't part of the original documents."'* Thus, audio characteristics (e.g., pitch and timbre) of text-to-speech voices could potentially convey the distinction between original and modified text content as well as which co-author acted on the content and how. However, participants emphasized the importance of attending to specific design choices that could facilitate (or even further complicate) the way screen reader users develop collaboration awareness. In particular, auditory representations that use multiple text-to-speech voices for collaborators' edits and comments must be designed carefully such that they do not create further cognitive overload for screen reader users instead of reducing it.

4.2 Understanding the Evolution of Document Content

In addition to understanding who edited what, an important aspect of collaboration awareness involves perceiving how original content changes through the underlying edits. For screen reader users, however, understanding the context of the edits becomes immensely difficult, because it requires them to keep track of the original text, edited text, and collaboration markup, all of which are intertwined in the screen reader read-aloud. Bill explained, *"What might take you 10 seconds to identify, may very well take me three minutes to disambiguate, because I'm going to read a complex paragraph with changes in complex sentences from three different authors, maybe even close to one another... I've forgotten the first half of the sentence by the time I get to the middle of the sentence."*

Instead of relying on screen reader announcements, sometimes Mike, Emma, and Henry keep multiple copies of a document – the original version and an edited version (without markup). They switch back and forth between these two copies and manually compare them sentence by sentence to detect *"how someone's changes would affect the document before and after."* However, this manual comparison process becomes quite challenging over time. As Mike described, *"...after a while, you can imagine what I decided to do. I quit. Because it didn't work."* Alternatively, sometimes participants ask their collaborators to summarize their edits using comments so that they can at least gather a high-level idea about *"what was in the original paragraph and what he has changed."* However, this workaround does not provide details about collaborators' actions that are visually available through features such as track changes or version history. *"You can imagine, it's not comprehensive enough compared to [what] you (sighted people) can see... very detailed changes that track changes can give you,"* commented Mike.

In this vein, participants emphasized the importance of listening to edits in the context of the original sentence so that they can easily figure out how collaborators' edits alter the form and meaning of the sentence. As Bill suggested, *"Read[ing] the original version versus the modified version would be super helpful and super powerful... to freeze the changes as if they have been accepted, and then to iterate across the possibilities... so that you can get a sense of what the different versions are."* These excerpts highlight the importance of a contextual presentation of edits that could make it easier for screen reader users to understand the evolution of text content through the edits.

4.3 Managing Disruption in Workflow

Our participants shared that the way screen readers provide notifications for collaboration markup through a series of spoken announcements often creates a *"verbal clutter."* Due to such continuous and copious collaboration notifications, participants find it extremely difficult to focus on their own work. Sofia shared her frustration: *"I was just hearing so much information that I just feel like I had a big jumbled mess in my document... It (track changes) didn't tell me enough of the information I needed, and it told me too much of the information I didn't need... I just didn't find it very effective for my workflow and my thought processing. It just made everything messier, not more efficient."*

Thus, in the process of making users aware of their collaborators' actions in the document, screen readers end up conveying *"too much information at once"* and impede individual workflow. To reduce such interruption in their workflow, participants often turn off spoken announcements and discuss their edits through alternative communication mediums instead (e.g., email, phone, or chat applications). Others often come up with workarounds to entirely avoid using default collaborative features. For instance, many participants prefer to *"read through this document and put any of my comments right in brackets or parentheses"* inline within the document text. Some also use special notations (e.g., @@@) that are *"unique enough that it's not going to be elsewhere in the document by chance"* so that they can easily search through the text for locating inline comments when needed.

Despite being a common strategy among our participants, leaving inline comments does not always work as the perfect solution either. Bill explained, *"It [inline comments] could be very helpful as you're just reading through maybe the first time you're getting a draft back from a colleague, but not as helpful if you're working on stuff and know about the comments already, and now they're all of a sudden getting in your way."* Here, we see that although inline comments are useful in certain scenarios, they can also be obtrusive to one's flow of reading and understanding of the text content in a similar way to spoken announcements.

Participants suggested one possible approach to reduce such verbal clutter and resulting disruption in workflow could be using non-speech audio cues (e.g., earcons) instead of spoken announcements to indicate the presence of comments or edits. Emma explained, *"I would like something less obtrusive... whether that be an audio cue or a notation on my braille display... Because the words (spoken announcements) are going to stop me from actually being able to listen to what I'm working on, where hopefully the not-words will not."* Prior work that focused on navigational tasks and perceiving auditory graphs have also found that non-speech audio cues are less disruptive and impose less cognitive load for processing information compared to speech [35, 70]. Importantly, screen reader users' preferences of auditory representations would likely be subjective and dependent on specific use-cases, where they *"might prefer speech in some instances, a tone [in others]."* Thus, a key design consideration for collaborative features is to determine which auditory representation works better for the task an individual is trying to accomplish at a particular instance.

4.4 Controlling the Influx of Collaboration Information

In addition to *how* collaboration information should be represented, *what* information needs to be accessed *when* also depends on the context of use, that is, whether the person is reading or editing the document at a particular instance and for what purposes. As such, understanding people's intent of use and customizing auditory representations accordingly are critical to *"siphon through"* the large amount of information that is generated in a collaborative writing scenario. Maya described, *"Collaboration is a good example where being able to customize the way information is presented will be really important, because different things will be important at different times. Maybe if I was an instructor, it would be really important for me to know that everyone's collaborating... But then maybe when I'm writing a paper, I really need to know the track changes that were added."*

In existing writing applications, screen reader users can control which collaboration information they want to know by toggling notifications for different collaborative features (e.g., comments, edits). While having separate controls for each feature can be useful in many scenarios, for screen reader users, *"that's just completely useless, because you have to remember to toggle all of those back, and they're not in the same place, and it's just an arduous task."* The way controls and settings options are designed in existing applications relies heavily on visual exploration and requires memorizing numerous keyboard shortcuts when accessed using screen readers. Bill instead suggested a mode or "scene-based" approach (e.g., editing or reading mode) that could allow people to easily consolidate and declare their desired collaboration information at a given instance. He said, *"What I would recommend there is a simple toggle of preference or verbosity, but not based around any type of static setting, but instead based around the fact that– 'Okay, I'm interested in a lot of editing related stuff now. Tell me about the following three, four, five things.'"*

Beyond mode-based controls, screen reader users also need to be able to control what information they hear at a particular instance by opportunistically navigating comments or edits as opposed to receiving continuous spoken notifications. Henry explained, *"Maybe you don't have to have all that information right away. If you just had an earcon or a beep sound, maybe that's- 'Hey, there's a comment here,' then you could press a hotkey to learn more about it."* As we see here, participants wanted to access collaboration information through a hierarchical approach that would combine

"less cluttered" notifications (e.g., non-speech audio cues) for the presence of a comment/edit with the opportunity to go deeper to explore the content of comments/edits with keyboard shortcuts.

Overall, our formative work illustrates that screen reader users face multiple challenges in maintaining collaboration awareness, which are due to the ways in which screen readers present who did what and where in a document as well as needing to understand how collaborators altered the document content without disrupting one's workflow. While there are several ways to address such challenges (e.g., text summarization [5]), our findings suggest one viable but relatively unexplored approach is to redesign the auditory representations that screen readers and word processing tools use to present collaborative information.

5 DESIGNING AND EVALUATING ACCESSIBLE AUDITORY REPRESENTATIONS

Building on insights from our formative interview study, we designed, developed, and evaluated auditory representations that aim to support collaboration awareness for screen reader users during asynchronous collaborative writing. The study investigates how different auditory representations can address issues of cognitive overload, verbal clutter, and lack of context associated with three key questions that are essential to developing collaboration awareness: (1) *where the comments are*, (2) *who commented what*, and (3) *who edited what*.

We conduct the study in three distinct modules, each centering around one of the three questions stated above. The auditory representations we developed focus on asynchronous collaboration information (e.g., comments and edits). In each module, We compare default techniques available in existing screen readers (i.e., direct spoken announcements) to one or two experimental auditory representations (i.e., non-speech audio and contextual presentation). Before conducting the study, we refined the auditory representations and our study design through in-person pilot testing sessions with three visually impaired writers. Pilot participants were expert users of screen readers (e.g., JAWS, NVDA, etc.) and familiar with collaborative features on Microsoft Word such as comments and track changes. The pilot sessions lasted for 60-80 minutes and participants received US$50 compensation.

Our formative study indicated that screen reader users' preferences regarding different auditory representations may depend on the context of use and complexity of the collaborative document. For example, a spoken announcement may seem overwhelming when multiple overlapping comments are attached within a sentence, whereas an earcon may work better in such a scenario. In contrast, if a comment is attached to a long span of text, a background tone alongside the text may give clearer indication of the presence of the comment, compared to earcons played only at the beginning and the end of the commented text. Thus, understanding the ways comment and edit complexity in a document influences the utility of different auditory representations is essential for making a collaborative system robust to the nuances that are likely to appear in natural writing situations. To address this, we designed the study to examine how participants' reactions to the default and experimental techniques are contingent upon collaboration complexity of the shared documents.

5.1 Generating Auditory Representations for Collaborative Writing

As the basis for our exploration of the various techniques, we developed a system that generates custom auditory representations using a Microsoft Word document as input. The system's input parameters can be configured to generate various audio representations corresponding to the document body text, edits, and comments. For example, these include adding earcons or additional contextual markers before or after comments or edits, playing edits and comments from different collaborators in different synthesized voices, and adding a background tone to indicate the presence of comments as the document content is read aloud. The system extracts text content and collaboration metadata from Word documents using the Word Object Model and pywin32 Python package.

It then converts these metadata to JSON objects and applies text-to-speech conversion and other non-speech audio effects to create different auditory representations. We used the Amazon Polly service for text-to-speech conversion and the LibROSA Python package for audio processing. We used a voice identified as male (Matthew) on Amazon Polly as the default voice that reads the main document content and collaboration markup phrases. Among all the English (US accent) voices available on Amazon Polly, Matthew was chosen as default, because it most closely matched with the default voices of JAWS and NVDA screen readers. We verified this with one of our pilot participants who is a proficient screen reader user. Full details of the specific auditory representations are provided in the three study module sections below (Sections 6.1, 7.1, and 8.1).

5.2 Experimental Design and Stimuli

We defined document complexity according to the key collaboration question that guides the design of each module, i.e., *where the comments are* in Module 1, *who commented what* in Module 2, and *who edited what* in Module 3. We employed a within-subjects design where each participant experiences all available techniques (default and experimental) in each module. For each technique, participants listen to two stimuli passages – one with high complexity and another with low complexity. Thus, each participant listens to 16 passages in total: six in the first module (two for each of the three techniques – one default and two experimental, see Section 6.2), four in the second module (two for each of the two techniques – one default and one experimental, see Section 7.2), and six in the third module (two for each of the three techniques – one default and two experimental, see Section 8.2). We prepared 16 different passages to ensure that participants do not experience a passage more than once. Each of these 16 passages had two variations accommodating two levels of document complexity. To save time during the study, we generated and pre-recorded audio for all the stimuli passages beforehand (see Section 5.1). Within each module, we fully counterbalanced the stimuli to control the presentation order of techniques and document complexity across participants.

We standardized the stimuli used in the study. Each stimulus includes a single passage (5-6 sentences with 45-55 words in the first module and 4-5 sentences with 35-45 words in the last two modules). All stimuli have readability scores ranging from 6-7 according to the Flesch-Kincaid Grade Level. While preparing the passages, we selected topics that were less likely to be of public knowledge but did not need domain expertise to be understood. We collected passages from online resources (e.g., Wikipedia, blogs) about birds, animals, cities, and landmarks, and cross-checked from multiple sources to ensure that the statements were factually correct. We chose commenter and editor names that were mono or bi-syllable (e.g., Lisa or Beth) and commonly used in English language. During our pilot sessions, we noticed that sometimes participants remembered comments/edits in terms of perceived gender identity of the voice (e.g., *"the boy made the most comments"*). While identifying a collaborator by the gender of their assigned voice could be useful in a natural writing scenario, it was potentially introducing a confound in our study. To avoid this, we included only female identifying names and voices for commenters and editors.

5.3 Participants

We conducted the evaluation study with 48 visually impaired writers who were randomly assigned to counterbalanced orders[2]. Eleven of our participants had also participated in the formative interviews. We recruited participants through the National Federation of the Blind, our research network, and snowball sampling. Each participant was compensated with a US$60 gift card.

[2]We were required to discard and re-run three sessions with new participants due to the presence of background noise and difficulty in understanding passage content.

Participants had different levels of visual abilities ranging from total blindness (60.4%) to legal blindness or low vision with or without light perception due to a number of conditions such as Retinitis Pigmentosa, Retinopathy of Prematurity, Glaucoma, etc., with onset at birth (60.4%) or later in life as well as acquired vision loss due to accidents. 47.9% of participants identified as female, 50% as male, and 2.1% as female/non-binary. Participants ranged in age from 19 to 60 with the most in the 25-34 range (37.5%). 62.5% participants identified as White, 16.7% as Hispanic, 10.4% as Asian, and 4.2% as Middle Eastern. Participant occupations included professor, assistive technology specialist, business analyst, finance advisor, attorney, and rehabilitation counselor, among others. Most participants (83.3%) lived in the U.S. and the rest came from seven different countries. Forty-two participants self-reported as expert users of one or more screen readers such as JAWS (45.8%), NVDA (37.5%), and VoiceOver (66.7%), while the remaining six participants self-reported as advanced users of at least one screen reader. Participants mostly used Microsoft Word (97.9%), Google Docs (91.7%), and text-based editors such as Notes or Notepad (93.8%) for writing. Many participants frequently used comments (60.4%), track changes (47.9%), and real-time editing (29.2%), while others used these features occasionally.

5.4 Procedure

We conducted the study remotely using the conferencing tool Zoom. We tested the audio quality on different networks and selected the best setup. We asked participants to work from a quiet space, with a reliable internet connection, and using the speaker configuration (headphones or speakers) that they prefer for working with screen readers. During the session, we played the audio stimuli on our local computer and shared computer audio with the participants via Zoom. Each evaluation session lasted for 80-100 minutes, was audio-recorded and later transcribed for analysis. All of the sessions were conducted by the first author between February 2020 and April 2020.

The study session started by explaining the purpose of the study and collecting verbal consent from the participants. For participants residing in EEA countries, we collected consent prior to the session using a GDPR-compliant online form. Next, we asked a number of questions related to participants' demographic information, usage of screen readers and writing tools, and their collaborative writing practices. After the demographic questionnaire, we played an example passage to adjust the volume level and check whether participants could hear the non-speech audio cues. We also requested that participants keep their screen readers muted (unless otherwise required), do not update their volume levels, and do not take any written notes during the session. Additionally, we explained to the participants that some questions may draw on their memory of the content presented and that they can respond with 'I don't recall the answer,' if needed.

We started each module by explaining the key collaboration question addressed in the module (e.g., 'where the comments are' in the first module) and the different techniques available to represent this information. At the beginning of each technique, we briefly explained how it works using an example passage and the kinds of questions participants will be asked after each passage. Participants had the option to listen to the example passage multiple times to understand the technique clearly. We used the same topic for the example passage throughout the study. Unlike the example passage, the stimuli passages were played only once during the main experiment.

After each passage, we asked a set of questions to assess participants' *perception of collaboration information* presented in the passage. We also asked one multiple-choice question specifically about the passage to gauge participants' *comprehension of passage content*. When participants finished listening to both passages for a technique, we asked them to rate their agreement with statements that captured their perception of use, i.e., *perceived ease of understanding collaboration information*, *perceived ease of learning*, *perceived cognitive load*, and *perceived disruption in workflow* on a 5-point Likert-style rating scale (ranging from 1-'strongly disagree' to 5-'strongly agree'). See Table 1 for

Table 1. **Study Measures**

	Module 1	**Module 2**	**Module 3**
Questions related to collaboration content and passage comprehension (asked after each passage)			
Perception of collaboration information	• Where are the comments attached– mostly in the first half, last half, or evenly distributed? • Are there any overlapping comments?	• Who commented about a specific text? • What did the commenter say about the specific text? • Who commented the most? • Are there any replies to a specific comment?	• Who edited a specific sentence? • How did the meaning of the sentence alter after the edits? • Who edited the most?
Comprehension of passage content	Example: where does the heron reside?	Example: where is blue lagoon located?	N/A
Likert-style self-report measures to capture participants' perception of use (asked after both passages for each technique)			
Perceived ease of understanding collaboration information	• I could easily understand where the comments were attached. • I could easily understand if there were any overlapping comments.	• I could easily understand who commented what. • I could easily understand what the comment was about. • I could easily understand the replies to a comment.	• I could easily understand who edited what. • I could easily understand how edits altered the meaning of a sentence.
Perceived ease of learning	This technique was easy to learn.	This technique was easy to learn.	This technique was easy to learn.
Perceived cognitive load	Understanding this technique required a lot of mental effort.	Understanding this technique required a lot of mental effort.	Understanding this technique required a lot of mental effort.
Perceived disruption in workflow	This technique disrupted my reading flow.	This technique disrupted my reading flow.	This technique disrupted my reading flow.
Open-ended questions about overall preference and further improvement (asked at the end of each module)			

the detailed study measures in each module[3]. We encouraged participants to rate these statements based on their overall experience with the techniques instead of whether or not they were able to answer passage comprehension and collaboration content related questions. We did this to reduce the extent to which participants' performance influenced their ratings, since the questions and the statements aimed to capture different facets of the techniques. Finally, at the end of the module, we asked participants open-ended questions regarding their *preferences* for different techniques (e.g., which technique(s) they liked the most and the least), the rationales behind their choices, and feedback for further improvement.

[3]We also administered questions and statements that capture perception of collaboration information in other nuanced aspects such as the span of commented text and the number of edits, comments, editors, or commenters in a passage. However, we do not see evidence of difference in the way different auditory techniques impacted these aspects. For ease of exposition, we do not report these results in the paper.

5.5 Analysis Method

We followed a mixed-method approach that involved quantitative analyses on performance measures and self-reported data as well as qualitative coding on open-ended feedback. Performance measures include responses to the questions we ask after each passage to capture participants' *perception of collaboration information* and *comprehension of passage content*. Self-reported data focus on participants' *perception of use* and are recorded as ratings regarding each technique as well as overall *preferences* for the techniques within each module.

For analyzing performance measures, two researchers independently reviewed and labeled participants' responses to each question with a binary category ('correct' and 'incorrect'). We assessed inter-rater reliability using Cohen's Kappa and achieved $\kappa = 0.83$ to $\kappa = 0.95$, which indicate high agreement among the coders [44]. We then resolved any disagreements through discussion. The predictor variables for our models depend on the specific module of the study under investigation. These included the *technique* experienced (e.g., default announcement, earcons, and tone overlay in the first module), *complexity* of the passage (low and high) and, when applicable, a *technique X complexity* interaction term. We also controlled for the *order* in which a participant experienced a technique (e.g., 1, 2, or 3) and a participant's *usage* of the relevant collaboration feature (i.e., whether they have prior experience of using the feature frequently or not).[4] We considered 'commenting' as the relevant feature in the first two modules that address where the comments are and who commented what. In the last module that addresses who edited what, we consider 'track changes' as the relevant collaboration feature. For analyzing the performance measures, we applied linear mixed effects logistic regression models to account for non-independence in the data (e.g., repeated measures collected from the same participants under different conditions).[5] For ease of exposition, throughout the paper, we only report results from the final models that were selected on the basis of Akaike Information Criterion (AIC) scores [25].

Similar to the models of performance measures, for self-reported ratings of perception of use, we included *technique* as a predictor and controlled for the *order* of experiencing the technique and a participant's *usage* of the relevant collaboration feature. We applied linear mixed effects regression models to analyze the self-reported ratings. Finally, we categorized participants' overall preferences on a scale of 1-3 (for the first and third modules) or 1-2 (for the second module) with higher rank associated with the most preferred technique. For analyzing these preference rankings, we included *technique* as a predictor and controlled for the *usage* of the commenting feature. We applied linear mixed effects regression models to analyze the preference rankings.[6] We report unstandardized β coefficients for linear regression throughout the paper, which permits interpretation of the predictor effects in original units.

Additionally, we analyzed participants' open-ended feedback using open-coding and iterative comparison between the codes to identify salient themes [22]. These concepts detailed participants' rationales behind preferring different techniques and how these techniques could improve and/or disrupt their perception of collaboration information and individual workflow in different contexts.

[4]We encountered model convergence issues in two cases, and removed control variables to address those, see Table 5 and Table 7.

[5]Linear mixed models have several advantages over Analysis of Variance (ANOVA) approaches including that they account for both fixed and random effects, and standard error adjustments are made to better account for repeated measures [37].

[6]We also applied a non-parametric statistical test (pairwise Wilcoxon Signed Rank test) on Likert-style responses and preference rankings and found nearly identical results.

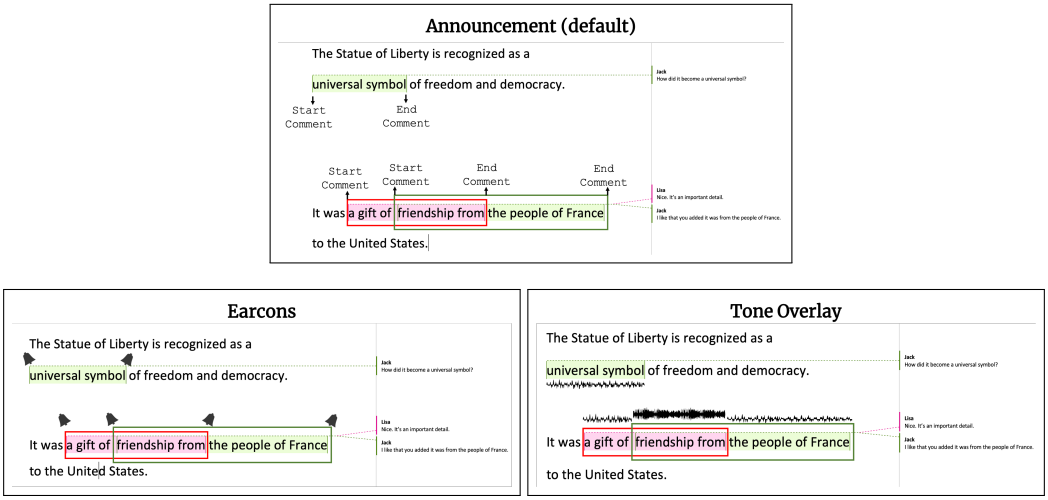

Fig. 2. Auditory representations in Module 1 (*where the comments are*) are shown for a passage with two sentences and three comments (second and third comments are overlapping). Top: Announcement. Collaboration markup is shown in `Courier New font`. Bottom left: Earcons. The bell icons slanted left and right respectively denote the starting DING sound and the ending DONG sound. Bottom right: Tone overlay. The low and high frequency waveforms respectively denote the lower and higher pitched background tone.

6 MODULE 1: WHERE ARE THE COMMENTS?

As the first component of our evaluation, we examine how to best support the challenging task of comprehending a passage while simultaneously identifying where comments are located and whether they are overlapping.

6.1 Auditory Representations

In the first module of the study, we incorporated three auditory representations to denote where comments are attached in a document: *announcement* (default), *earcons*, and *tone overlay*. Following insights from our interview study, we designed the earcons and tone overlay representations to assess whether non-speech audio can reduce the 'verbal clutter' created by the spoken announcement while indicating the location of comments in a document.

Announcement (default). Spoken announcement is the default technique that many screen readers use to indicate the presence of a comment attached to a text portion of the document content. Different screen readers use slightly different phrases to announce the starting and ending of a comment. We chose the phrases '`start comment`' and '`end comment`' following the way JAWS announces comments in Google Docs. In cases where two comments overlap each other, two '`start comment`' phrases appear one after another, indicating that a comment started before another comment ended (i.e., it was fully or partially overlapped, see Figure 2, top). Note that screen readers use a variety of speech-based configurations to represent comments and edits. For example, both JAWS and NVDA have list views where users can navigate through all the comments (or edits) sequentially. However, we chose the aforementioned technique as the default, since our focus in this study is on the way screen reader users consume collaboration information as they go through the document content – not on how they attend to the list of comments/edits separately. An audio example can be found here: http://bit.ly/bvi-cw-mod1-announcement.

Earcons. In this technique, two distinct audio tones work as earcons [35] i.e., abstract representations of the spoken phrases '`start comment`' and '`end comment`'. We used a two-part bell sound (DING-DONG), where the DING sound specifies the starting of a comment and the DONG sound specifies the ending (see Figure 2, bottom left). Similar to the announcement technique, two DING sounds appearing one after another indicates the overlap between two comments. Based on feedback received from pilot testing sessions, we adjusted the length and the loudness of the sounds to make them noticeable but subtle and comfortable for listening. For the same reason, we chose these short-lived DING-DONG sounds instead of complex earcons consisting of multiple rhythmic sequences [35]. An audio example can be found at this link: http://bit.ly/bvi-cw-mod1-earcons.

Tone Overlay. In this technique, a tone is continuously played in the background as long as the text portion associated with a comment is read out. The frequency (i.e., pitch) of the background tone is increased when text portions have multiple comments overlapping with each other so that users can detect where standalone and overlapping comments are attached (see Figure 2, bottom right). We used 185 Hz (note G3) for the background tone associated with text having standalone comments and 220 Hz (note A3) for overlapping comments. We adjusted the amplitude of the background tone according to feedback from pilot participants to keep it at a discernible level but much lower than the level of the text read-aloud. We did this to ensure that users can distinguish the background tone from the text read-aloud, but it does not impede perception of the text content. An audio example can be found at this link: http://bit.ly/bvi-cw-mod1-tone-overlay.

6.2 Stimuli and Measures

Given the focus on understanding where the comments are attached, we manipulate document complexity in terms of the number of comments, the length of the text where comments are attached, and whether there are any overlapping comments. For each of the three techniques, we prepared stimuli passages with two levels of complexity: low (2-3 comments in total, all are attached to 2-4 words in the passage text, and no overlapping comments) and high (5-6 comments in total, two of them are attached to a single word, one with a whole sentence and the rest to 2-4 words, and one pair of overlapping comments). Thus, in this module, participants listen to six passages in total, two for each technique. Table 1 includes the set of questions we asked to assess participants' *perception of collaboration information* and *comprehension of passage content* and the self-report statements we administered to capture their *perception of use* and overall *preference*.

6.3 Results

We begin by investigating how different auditory representations (default announcement, earcons, or tone overlay) affect participants' performance on the questions related to where the comments are attached. With regards to the question that asked about the location of the distribution of comments in the passage (i.e., whether the comments are attached mostly in the first half or last half of the passage or almost evenly distributed throughout), we see differences in the way participants performed using these techniques in low and high complexity passages. Specifically, using earcons, participants were less likely to correctly identify the location of the comments in a low complexity passage relative to the default announcement, whereas they were more likely to correctly identify comment locations in a high complexity passage relative to the default announcement. In other words, the odds of correctly locating comments with earcons is 0.33 times than with the default announcement in a low complexity passage, whereas in a high complexity passage, the odds of correctly locating comments with earcons is 6.4 times than with the default announcement (for the interaction, $log(OR) = 2.95$, $p = 0.006$, see Figure 3 and Table 3). There was a similar statistical trend in participants' performance using tone overlay in low and high complexity passages. Particularly,

in a low complexity passage, the odds of correctly identifying the location of comments is 0.57 times compared to the default announcement, whereas in a high complexity passage the odds of correctly identifying the location of comments is 4.1 times compared to the default announcement (for the interaction, $log(OR)$ = 1.96, p = 0.059, see Figure 3 and Table 3). This possibly indicates that in a low complexity passage with only a few comments dispersed throughout, the spoken announcements may have provided a more straightforward way to understand where comments are located compared to non-speech audio cues. However, in a high complexity passage where comments were densely populated — in close proximity and with overlaps between each other, spoken announcements may have become more confusing and verbose while earcons and tone overlay performed relatively better in identifying the distribution of comments.

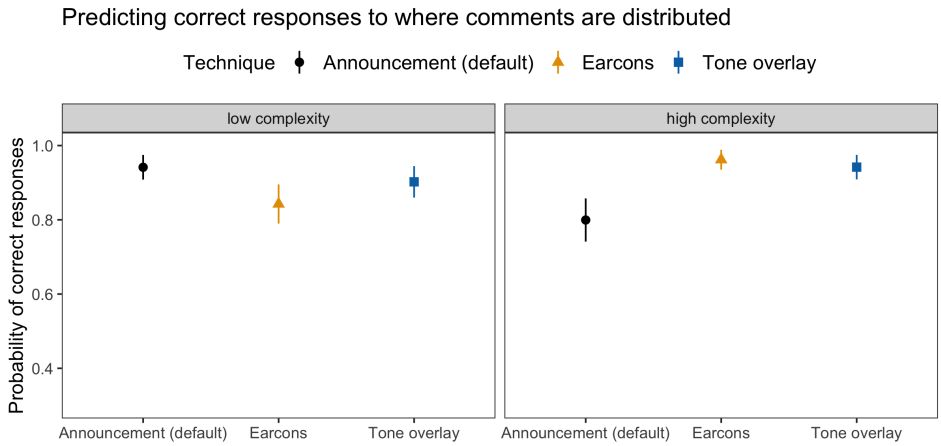

Fig. 3. Plot showing predicted values for correct responses to the question about the location of the distribution of comments using default announcement, earcons, and tone overlay in low and high complexity passages. Error bars represent +/-SE.

Turning to the comprehension of passage content, we see that tone overlay improved comprehension relative to the default technique and earcons. With tone overlay, the odds of correctly answering the question about passage content is 1.9 times compared to using the default technique ($log(OR)$ = 0.63, p = 0.055, Table 3) and 2.7 times compared to using earcons ($log(OR)$ = 1.01, p = 0.002)[7]. In addition, participants reported several benefits of tone overlay and earcons on the self-report measures. They felt it was easier to understand overlapping comments using tone overlay compared to the default technique and earcons; particularly, the predicted rating for tone overlay is 0.58 units higher (on the five point Likert-scale) than for the default technique (β = 0.58, p = <0.001, Table 4) and 0.31 units higher than for earcons (β = 0.31, p = 0.046). Additionally, they reported that their reading flow was less disrupted using both earcons (β = -0.50, p = 0.04, Table 4) and tone overlay (β = -1.02, p < 0.001, Table 4) compared to the default announcement. This finding that both earcons and tone overlay were considered less disruptive than spoken announcement supports our intuition behind using these non-speech audio representations to reduce verbal clutter and help fluent reading flow. Furthermore, tone overlay was considered to be even less disruptive (β = -0.52, p = 0.03) and requiring less cognitive effort (β = -0.44, p = 0.045) than earcons.

[7]To directly compare earcons and tone overlay, we re-ran the models and changed the reference category of the *Technique* variable from default announcement to earcons. For this reason, these results do not appear in the table in Appendix B.

Overall, these results illustrate that non-speech audio such as tone overlay better represented some aspects of collaboration information (e.g., overlapping comments) without creating much disruption in the reading flow and were not detectably better or worse than the default announcement in other aspects. This is also supported by participants' overall preference for the techniques. Although there was no significant difference between earcons and the default announcement (β = 0.22, p = 0.17, Table 4), participants preferred tone overlay more than the default announcement (β = 0.63, $p < 0.001$, Table 4) and earcons (β = 0.41, p = 0.01).

Our qualitative analyses of participants' open-ended feedback provided deeper insights into how these auditory techniques supported and impeded their understanding of the passage content and the presence of comments. One factor that considerably influenced participants' preferences was to what extent a technique helped them disambiguate text content and collaboration information. Many participants (52.1%) preferred tone overlay, because it uses *"verbal and non-verbal cues, so it is easier to distinguish the text"* (P41) and they *"could kind of visualize words being underlined or highlighted"* (P37). Participants' reactions also depended on the way non-speech audio cues shifted their attention from text content to collaboration information. For instance, some felt that with earcons and tone overlay, they *"ended up paying attention more to the tone than the audio (speech)"* (P13). This reaction may have stemmed from the fact that our experimental techniques such as tone overlay were novel to many participants and they thought *"it would take a little bit longer to get used to it"* (P23). Participants also added that non-speech audio cues need to be customizable according to one's individual receptivity towards audio enhancements and hearing abilities.

Additionally, participants found tone overlay to be more helpful in perceiving the span of commented text: *"tone [overlay] was continuous through the comments, so you are kind of aware that you're still in the comment versus not in a comment"* (P9). However, compared to earcons, tone overlay was *"a little less precise in terms of where the comment starts and ends"* (P8). To get the benefits of both earcons and tone overlay, participants recommended combining these two techniques or choosing one based on the span of the commented text: *"maybe have a ding and a dong (earcons) for a single word but a tone [overlay] for a sentence"* (P34).

The amount of time required to perceive collaboration information was another key consideration for our participants. They felt that earcons were *"quicker to read"* (P19) and *"fleeting"* (P44) compared to spoken announcements. Tone overlay was even better, *"because you are getting two pieces of information at once... it will represent a huge productivity boost. You are reading the text and you are getting an immediate indication that that text is commented"* (P25).

While a majority of the participants preferred some form of non-speech audio, those who preferred the default announcement mentioned familiarity as a key reason. P14, an IT professional, preferred the default announcement *"probably because it's similar to other materials that I've read that also have similar tags like HTML or different object notation things in programming that indicate the beginning and ending of particular blocks."* Participants who preferred non-speech audio mentioned additional concerns, such as memorizing *"too many other sounds that were used to indicate the beginning and ending of things... that's a little harder to keep track of which sounds are which things"* (P8). Prior work has also highlighted that earcons require explicit learning [35]. To address this, P7 and P17 suggested using easily distinguishable earcons that can be meaningfully mapped to the notion of opening and closing comments, such as *"the train tones... like opening and shutting doors."*

In summary, our quantitative and qualitative results suggest that: (1) tone overlay best supports the challenging task of identifying where comments are located without causing disruption to reading flow; (2) earcons and tone overlay are most useful in understanding where comments are located in complex passages (i.e., densely populated with comments); and (3) these techniques may work best in combination depending on the document complexity (e.g., presence of overlapping comments and the span of commented texts).

7 MODULE 2: WHO COMMENTED WHAT?

In addition to understanding where comments are located, writers must also understand the content of the comments as well as who amongst multiple collaborators made the comment–all without disrupting their ability to comprehend the document content.

7.1 Auditory Representations

The second module of our study incorporates two different techniques to present the content of comments (or replies) and the name of the commenters: *reading inline with a consistent voice* (default) and *reading inline with voice coding*. The location of a comment in the passage is indicated by spoken announcement used by many screen readers, i.e., '`start comment`' (see the default technique used in Module 1, Section 6.1).

Reading Inline with a Consistent Voice (default). In this technique, a comment (or reply) and the name of the commenter are read out in-line with the main text just after reading the text portion where the comment is attached. If there are replies associated with the comment, those are also read out sequentially along with the names of the replier. After reading the comment (and replies), it goes back to read the rest of the passage from where it left off. To distinguish between the end of a comment and the rest of the passage, it announces a signposting message '`back to document`' (see Figure 4, left). Screen readers typically do not read comments in-line with the main text. However, we consider this as the default technique for presenting comment content, since most participants in our formative study used inline comments and preferred this technique over the traditional commenting feature for sharing feedback with their collaborators. An audio example can be found at this link: http://bit.ly/bvi-cw-mod2-same-voice.

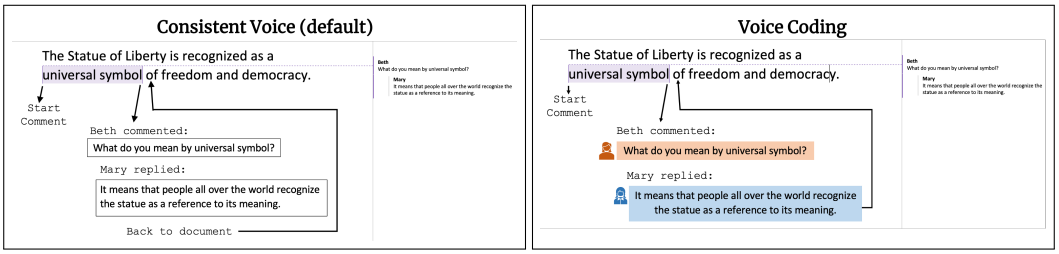

Fig. 4. Auditory representations in Module 2 (*who commented what*) are shown for a sentence with a comment and a reply from two individuals. Collaboration markup is shown in `Courier New font`. Left: Reading inline in a consistent voice. Right: Reading inline with voice coding. The comment and reply highlighted in different colors (orange and blue) are read by different voices. Note that the text 'universal symbol' is read in the default voice; it was highlighted by Word application since a comment was attached to it.

Reading Inline with Voice Coding. This technique applies voice coding on commenters, i.e., it assigns specific text-to-speech voices to commenters on a document. The content of the comments (and replies) left by an individual is read out in the voice associated with them (see Figure 4, right). Building on insights from our formative findings, this technique incorporates multiple text-to-speech voices to disambiguate between document content and contributions by different collaborators. An audio example can be found at this link: http://bit.ly/bvi-cw-mod2-voice-coding.

7.2 Stimuli and Measures

In this module, which focuses on understanding who commented what, we manipulate document complexity in terms of the number of commenters and whether a comment has any replies or not.

Similar to the previous module (Section 6.2), we prepared passages with two levels of complexity: low (two commenters leaving four comments, no replies) and high (four commenters leaving four comments and two replies). Thus, in this module, each participant experiences four passages in total, two for each technique. In each passage, one commenter makes a higher number of comments than the rest, where the difference between two individuals' comments are at least two. We did this to ensure that each passage has a salient contribution from one individual so that we can assess whether any of the audio representations perform better in making this distinction clearer than the rest. Table 1 includes the set of questions we asked to assess participants' *perception of collaboration information* and *comprehension of passage content* and the self-report statements we administered to capture their *perception of use* and overall *preference*.

7.3 Results

In this module, we first investigate whether reading comments using voice coding improves participants' understanding of who commented what in a passage in comparison to the default technique that reads comments in a consistent voice. In response to the questions that asked who commented about a specific text, participants were more likely to provide correct answers using voice coding compared to the default technique. Specifically, we see that the odds of providing a correct answer with voice coding is 4.2 times relative to the default technique ($log(OR)$ = 1.43, p < 0.001, Table 5). We see a similar pattern in participants' responses to the question about what the commenter said in the specific comment. In this case, the odds of providing a correct answer with voice coding is 2.8 times relative to the default technique ($log(OR)$ = 1.03, p = 0.002, Table 5). These results suggest that multiple text-to-speech voices may have helped participants better disambiguate between contributions by different commenters.

While the previous results suggest an encouraging effect of voice coding on participants' perception of who commented what, we see a different outcome regarding their comprehension of passage content. Participants were less likely to answer correctly to the question about passage content using voice coding compared to the default technique. With voice coding, the odds of providing a correct answer is 0.45 times relative to the default technique with a consistent voice ($log(OR)$ = -0.81, p = 0.03, Table 5). One possible explanation for this could be that because of the different voices associated with comments, participants may have paid more attention to the comments instead of the passage content, and as such, they may not have been able to correctly answer the passage related questions.

In response to the question about who made the most comments, participants were more likely to correctly answer using voice coding compared to the default technique in a low complexity passage. However, in a high complexity passage, they were less likely to provide correct answers to this question using voice coding compared to the default technique. Specifically, with voice coding, the odds of correctly identifying who commented the most is 1.8 times than with the default technique in a low complexity passage, whereas in a high complexity passage, the odds of correctly answering this question with voice coding is 0.3 times than with the default technique (for the interaction, $log(OR)$ = -1.77, p = 0.02, see Figure 5 and Table 5). This result may have occurred because high complexity passages contain a higher number of commenters. Thus, the increased number of distinct text-to-speech voices associated with these commenters with voice coding may have made it difficult to keep track of which voice read the highest number of comments.

Turning to participants' self-report measures, we see that participants found replies to a comment easier to understand with voice coding than the default technique: particularly, the predicted rating for voice coding is 0.19 units higher (on the five point Likert-scale) than that for the default technique (β = 0.19, p = 0.02, Table 6). Participants also found voice coding to be less disruptive than the

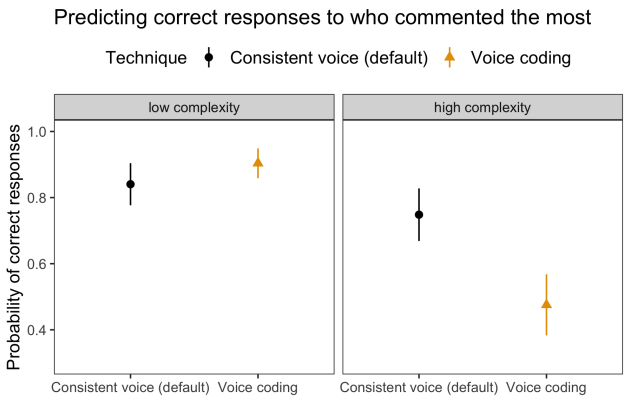

Fig. 5. Plot showing predicted values for correct responses to the question about who commented the most using a consistent voice and voice coding in low and high complexity passages. Error bars represent +/-SE.

default technique (β = -0.42, p = 0.03, Table 6). Overall, participants preferred voice coding to the default technique (β = 0.60, p < 0.001, Table 6).

To further understand participants' reactions to these techniques, we take a closer look at their open-ended responses. As we expected, the most important factor that contributed to many participants' (79.2%) preference for voice-coding is that having different voices *"really makes a contrast between the actual passage and the comments and who replied"* (P13). Participants further added that voice coding helped them *"actually visualize [commenters] having a conversation, so it was much more animated, much more tangible, much more concrete"* (P36).

Some participants, however, showed an opposite reaction towards voice coding, particularly regarding the extent to which it shifted their attention away from text content, as we see happening with non-speech audio cues in Section 6.3. This also aligns with our quantitative results where we found that voice coding negatively impacted participants' comprehension of passage content. Relatedly, participants shared concern regarding the amount of cognitive effort required to keep track of collaboration information elicited by different voices. Particularly when a large number of commenters contribute to a document, participants felt that it might be *"a nightmare"* (P1) to figure out what voice corresponds to whom. This provides a possible explanation of why voice coding was less helpful in detecting who commented the most in a high complexity passage.

To address this, participants suggested using *"just one voice for all commenters... as opposed to different voices for each person"* (P4) so that they can differentiate between text content and comments but do not get overwhelmed with excessive auditory clutter. Additionally, some participants recommended using easily distinguishable and representative voices for reading comments, such as *"accented"* voices that *"sounded a bit like them (co-authors)"* (P1). Having such personalized voices could also help people get rid of extraneous spoken announcements, because *"you don't have to read the [commenter's] name every time... when I can associate a voice to it"* (P23). Even participants who did not prefer listening to entirely different voices recommended other non-speech audio enhancements to distinguish between document content and comments, such as changing the pitch or spatial location of the default voice or adding a background sound while reading comments (similar to the tone overlay technique used in Module 1, Section 6.1).

In summary, our results indicate that voice coding (1) improved participants' ability to identify which collaborator commented about a specific text and what they said in their comment; but, voice coding also (2) posed a distraction when attempting to comprehend a passage and was problematic when used to identify who commented the most in a passage with more than two commenters.

8 MODULE 3: WHO EDITED WHAT?

Not only must writers keep track of where comments are located, what information each comment contains, and who added various comments, they must also understand which collaborators made certain edits and how those edits changed the document.

8.1 Auditory Representations

In the third module of our study, we incorporated three auditory representations for describing who edited what in a document: *announcement* (default), *contextual presentation*, and *contextual presentation with voice coding*.

Announcement (default). This technique announces the name of the editor and the type and content of an edit while reading the document text using spoken phrases, such as 'Mary inserted' and 'end insertion'. For example, consider the sentence in Figure 6, top, that has an insertion and a deletion from two editors. This sentence is read by a screen reader as follows:

> "`The statue of liberty was a gift <pause> Mary inserted <pause> of friendship from the people of France <pause> end insertion <pause> Beth deleted <pause> the people of <pause> end deletion <pause> to the United States.`"

We consider this technique as the default (see Figure 6, bottom left), since it aligns with the way many screen readers describe edits made with the Track Changes feature on a Word document. An audio example can be found here: http://bit.ly/bvi-cw-mod3-announcement.

Fig. 6. Auditory representations in Module 3 (*who edited what*) are shown for a sentence with two edits from two individuals. Collaboration markup is shown in `Courier New font`. Top: Sample sentence as seen with 'All Markup' option on MS Word. Mary inserted the underlined text ('of friendship from the people of France') and Beth deleted the text marked up with strikethrough ('the people of'). Bottom left: Announcement (default) technique. Bottom middle: Contextual presentation. Bottom right: Contextual presentation with voice-coding. Text highlighted in blue is read by the voice assigned to the corresponding editor, Mary.

Contextual Presentation. This technique reads a document sentence-by-sentence presenting edits in-context of the sentence. That is, to contextually present a suggested edit, it reads the corresponding sentence as it would have appeared after the edit was applied to it. To make the effect of the edit more salient, this technique presents both versions of the sentence– before and after the edit is applied. It first starts with reading the original version before any edits are applied, followed by announcing the number of edits in the sentence and reading different versions of

the sentence after applying the suggested edits sequentially one after another. Returning to the example in Figure 3 (top), Mary's edit (i.e., inserted text) occurred earlier than Beth's edit (i.e., deleted text). These edits are presented sequentially (See Figure 6, bottom middle):

> "`The Statue of Liberty was a gift from France to the United States <pause> Two edits <pause> Insertion by Mary <pause> The Statue of Liberty was a gift of friendship from the people of France to the United States <pause> Deletion by Beth <pause> The Statue of Liberty was a gift of friendship from France to the United States.`"

Building on the insights gathered from our formative findings, this technique highlights how edits alter the meaning of a sentence by presenting those edits within the context of the sentence. An audio example can be found at this link: http://bit.ly/bvi-cw-mod3-contextual.

As shown in the example above, while reading a version of the sentence corresponding to a specific edit, this technique retains the edits that were made earlier. We do this considering the possible interdependence between sequential edits (e.g., an editor may delete a word that was previously inserted by another editor, thus the deletion cannot be understood without the earlier insertion). Additionally, sequential presentation highlights the way a sentence evolves throughout the course of the suggested edits. However, it is also important to understand how an individual edit can alter the meaning of the sentence. Following the same technique we used, it is possible to iterate through all possible versions of a sentence applying the suggested edits individually as well as in combination with other edits.

Contextual Presentation with Voice Coding. This technique is a variation of the contextual presentation technique, where text portions inserted by different editors are voice coded, i.e., read out in the editors' respective synthesized voices. In the previous example, the text portion inserted by Mary ('`of friendship from the people of France`') is read out in the synthesized voice assigned to them. Collaboration markup phrases (e.g., '`Insertion by Mary`') and text portions written without Track Changes are read in the default voice (see Figure 6, bottom right). Similar to the contextual presentation technique without voice-coding, this technique retains earlier edits in each iteration of a sentence that has multiple edits. To address the concern about cognitive overload in listening to several different voices within a sentence, we refined the technique to read the earlier edits in the default voice, while only the text portion inserted in the current iteration is read in the editor-specific voice. This limits the number of distinct voices in each iteration of a sentence to two: the default voice and the one associated with the editor who made the edit in the current iteration. In this way, it also highlights the content of the current edit by making it stand out amidst text read out in the default voice. An audio example can be found at this link: http://bit.ly/bvi-cw-mod3-contextual-voice-coding.

8.2 Stimuli and Measures

While the previous two modules focused on the comments in a passage, this module examines tracked changes or edits on the passage content. We manipulate document complexity in terms of the number of edits, editors, and overlapping edits (i.e., one editor deleting a word from a text portion that another editor inserted). We prepared passages with two levels of complexity: low (two editors, four edits in total, no overlapping edits) and high (four editors, six edits in total with two pairs of overlapping edits). Thus, in this module, each participant experiences six passages in total, two for each technique. Similar to Section 7.2, we ensure that each passage has a salient contribution from one individual in that they make a higher number of edits (at least by two) than the rest of the editors. Table 1 includes the set of questions we asked to assess participants'

perception of collaboration information and the self-report statements we administered to capture their *perception of use* and overall *preference*. However, unlike previous modules, we do not ask any questions about the passage content separately, since the question about the changes in the meaning of a sentence after suggested edits also captures comprehension of passage content.

8.3 Results

We start by analyzing whether contextual presentation and contextual voice coding affect participants' perception of who edited what differently than the default announcement. Looking at the responses to the question about who edited a sentence, we see that participants were more likely to correctly answer using contextual presentation relative to the default announcement. Specifically, the odds of providing a correct answer is 3.1 times in the contextual presentation relative to the default technique ($log(OR) = 1.12$, $p = 0.007$, Table 7). We see an even larger effect in response to this question with contextual voice coding: the odds of providing correct answers is 4.4 times in contextual voice coding relative to the default announcement ($log(OR) = 1.48$, $p < 0.001$, Table 7). This indicates that the addition of voice coding with the contextual presentation may have been more helpful in recognizing the editor correctly.

Similarly, with regards to the question about how the edits altered the meaning of a sentence, participants were more likely to provide correct answers using both contextual presentation and contextual voice coding techniques compared to the default announcement. We see that the odds of providing correct answers is 5.6 times in contextual presentation compared to the default technique ($log(OR) = 1.73$, $p < 0.001$, Table 7). Contextual voice coding shows a similar pattern with an even larger effect: the odds of providing correct answers is 15.4 times compared to the default technique ($log(OR) = 2.73$, $p < 0.001$, Table 7). Further, participants were more likely to provide a correct answer to this question in contextual voice coding relative to the contextual presentation technique. The odds of providing correct answers is 2.7 times in contextual voice coding compared to the contextual technique ($log(OR) = 1.01$, $p = 0.006$). This indicates that including voice coding in the contextual presentation may have helped participants identify the newly inserted text in the modified version of the sentence and thus provided an even better understanding of how the meaning of the original sentence changed.

Participants' self-reported ratings bolstered the results reported above. In particular, understanding who edited what was perceived to be easier using contextual voice coding compared to the default technique: the predicted rating for contextual voice coding is 0.38 units higher (on the five point Likert-scale) than the default technique ($\beta = 0.38$, $p = 0.03$, Table 8). Similarly, participants found it easier to understand changes in the meaning of a sentence with contextual voice coding than the default technique ($\beta = 0.58$, $p = 0.002$, Table 8) and contextual presentation ($\beta = 0.35$, $p = 0.057$). This result aligns with our formative findings, which inspired us to present collaborators' edits in-context by using different voices to iteratively highlight how the edits alter the meaning of the original content.

Additionally, compared to the default announcement, participants found contextual voice coding easier to learn ($\beta = 0.42$, $p = 0.02$, Table 8), requiring a lower cognitive load ($\beta = -0.73$, $p < 0.001$, Table 8) and causing less disruption in reading flow ($\beta = -0.83$, $p < 0.001$, Table 8). Further, contextual voice coding was rated as less disruptive ($\beta = -0.54$, $p = 0.003$) and requiring less cognitive effort ($\beta = -0.42$, $p = 0.02$) than contextual presentation. When we look at participants' overall preferences, we see that contextual presentation was preferred to the default technique ($\beta = 0.32$, $p = 0.04$, Table 8) and contextual voice coding was preferred to both the default technique ($\beta = 0.71$, $p < 0.001$, Table 8) and contextual presentation ($\beta = 0.39$, $p = 0.01$). Overall, these results illustrate that contextual presentation technique improves perception of edits, and the integration of voice coding with this technique makes it even better by reducing cognitive load and disruption in workflow.

Participants' open-ended responses further strengthened our findings from quantitative analyses. The key reason that guided most participants' preference for contextual voice coding is that this technique combined the benefits of presenting edits in-context of the original sentence along with the *"extra reinforcement"* by different voices that served as *"an easier memory guide for who did what and also what was [an] edit and what wasn't"* (P22). In contrast, the default spoken announcement, which was the least preferred by most participants (60.4%), was considered as *"a complete waste of time"* (P44), because participants felt that they *"couldn't really get a good grasp of how it changed the meaning by listening through it just a one time... I will have to go through it a few times"* (P23).

Despite important benefits of contextual voice coding, the different voices caused a distraction for some participants, as we also found in Module 2 (Section 7.3). P18 further explained that the natural break that occurs in the synthesized speech when a new voice with different prosody appears in the middle of a sentence *"actually destroys the intonation of the sentence. So, a screen-reader user who is expecting a sentence to come in a natural flow loses that track."* Instead of assigning distinct voices to individual editors, these participants suggested having a single voice to read all the edits, as they did in the previous module (Section 7.3). Relatedly, those who preferred the default technique due to its simplicity, still wanted other forms of non-speech audio such as earcons, tone overlay, and changing pitch or voice of spoken announcement phrases to distinguish edits from text content. Interestingly, as we also found in Module 1 (Section 6.3), participants considered tone overlay to be better than earcons because of its *"efficiency"* in terms of time requirement and clarity in depicting the span of the edited (or commented) text.

While contextual presentations (with or without voice coding) were generally preferred to the default technique because of the reasons discussed above, many participants expressed concern about the repetition of a sentence in contextual presentation– the reason why it takes longer to finish a passage in this technique. P1 said, *"As blind people, things generally take us longer and every time the sentence is read a second time, I'm like- 'okay, I already heard that', and if you're talking [about] a long document, that's gonna take an age to go through."* As such, some participants said that in a natural writing scenario, they would prefer listening to only *"the focus or the area that was changed, either just the words that were added or deleted, or maybe the immediate context"* (P4) instead of the original and modified versions of the entire sentence. Participants also emphasized that interactive collaborative features that would allow them to consume information on an as-required basis might further reduce disruption in their workflow and improve collaboration awareness.

In summary, our quantitative and qualitative results show that (1) the combination of contextual presentation and voice coding provides the best support for understanding who edited a sentence and how the edits altered the meaning of a sentence; however, (2) presenting edits in the context of an entire sentence requires more time than the default technique, and (3) changing voices in the middle of a sentence to present edited text can break the continuity of reading.

9 DISCUSSION

Maintaining collaboration awareness is a complex challenge for all writers. Yet, the serial nature of how screen readers present text-based content, combined with the lack of well-designed auditory representations for collaborative markup makes the work of achieving and maintaining collaboration awareness particularly difficult for blind writers. Prior research in HCI has discussed the problematic relationship between accessibility and usability [11, 47, 65, 78, 81], showing that many technological systems are accessible on the surface but not usable for practical purposes [11, 29]. Our analysis reveals a specific instance of this problem: screen reader users have difficulty not only developing collaboration awareness but also maintaining efficiency due to tools that are *"supposedly accessible but very poorly implemented"* [29]. As such, writing tools must be designed such that screen reader users can perceive collaborative information efficiently without additional

cognitive effort or significant disruption to their individual workflow (e.g., reading or writing on their own). The present paper provides a foundation for creating more accessible collaborative writing tools through our empirically grounded design and evaluation of multiple auditory techniques for asynchronous collaborative writing. Below, we discuss the design tradeoffs and considerations for auditory representations that address issues of cognitive effort, disruption, and efficiency.

9.1 Managing Cognitive Effort in Understanding Collaboration Information

Our formative study revealed that screen reader users need to apply a higher amount of cognitive effort in sifting through the *"jumbled mess"* of collaboration notifications that appear in the same format as the one used to read text content, i.e., speech. In contrast, by presenting collaboration information in a distinct auditory format, non-speech audio cues and voice coding help people perceive the location, content, and author of comments and edits with less cognitive effort. Specifically, voice coding makes it easier to keep track of who commented or edited what, while non-speech audio cues (e.g., tone overlay) are helpful in distinguishing between text content with overlapping or standalone comments, or without any comments attached. Interestingly, some of these techniques helped our participants create a mental imagery of collaborators' actions. For example, tone overlay worked as an auditory *"underline or highlight,"* while voice coding created an impression of people *"having a conversation."* Thus, the auditory enhancement and expressiveness non-speech audio and voice coding offer can minimize the cognitive effort [70] required to disambiguate between complex and intertwined pieces of collaboration information and text content.

Despite this benefit, mapping non-speech audio cues to their corresponding meanings [35] or figuring out which voice refers to whom can put additional cognitive load on screen reader users, particularly when various pieces of information (e.g., starting and ending of comments, insertions, deletions, etc.) are indicated by non-speech audio cues or a large number of co-authors contribute to the shared document. In contrast, spoken announcements that provide straightforward description of the collaboration markup (e.g., 'start comment', 'end comment') do not require explicit semantic mapping or memorizing. As such, spoken announcements may be preferable for novice screen reader users when they are just starting to use collaborative tools, whereas people may switch over to non-speech audio and voice coding techniques when they have a better understanding of the syntax of collaborative features and semantic mappings of audio cues. Another approach to address this issue could involve using representative auditory icons [35] (e.g., the sound of a door opening or closing) instead of abstract earcons. Furthermore, collaborative tools and screen readers could allow users to create personalized voice profiles for their co-authors [1]. This could potentially reduce the cognitive load of mapping different voices to co-authors, especially when working with the same collaborators (e.g., manager or advisor) and the voices become familiar over time.

Our analysis also illustrated that different auditory techniques can incur more or less cognitive load depending on the specific collaboration information they are presenting and the level of collaboration complexity in the document. For example, earcons can point to the precise locations where a comment starts and ends, whereas tone overlay can provide a clear understanding of the span of a comment and where comments overlap with each other. As such, audio representations should be implemented in a way that aligns with the context in which they are being used [36] (e.g., a complex document with large number of edits or a paragraph with overlapping comments) and may work best in combination, i.e., screen readers could dynamically render collaborative information based on the complexity and structure.

9.2 Reducing Disruption in Individual Workflow

Our analysis joins prior work in highlighting the ways screen reader representations pale in comparison to the mainstream collaborative features that are designed for sighted people [9, 67, 84].

One example of this is the way collaborative tools leverage glanceability [67] to present multiple layers of collaboration information in tandem with text content through color-coding and comment sidebars, whereby sighted people can direct their attention to where they want to focus on by a quick glance without interrupting their current task at hand. In contrast, screen readers push spoken alerts to describe collaboration information interlaced with text content, creating a continuous disruption to one's own reading flow. In this regard, non-speech audio cues (e.g., earcons and tone overlay) can offer a *"less obtrusive"* approach for making sense of collaboration information.

While non-speech audio cues are generally less disruptive than spoken announcements, our analysis showed that audio cues can also sometimes pose a distraction and shift people's attention away from understanding text content. Similarly, changing voice in the middle of a sentence can break the continuity of intonation and prosody in a way that may become *"jarring"* and *"discordant."* To address this, participants wanted to have a single voice for all commenters (or editors) that will be distinct from the default voice for text content. Thus, a simpler version of voice-coding (or manipulation of pitch or timbre) could make it easier to differentiate between text content and comments/edits without breaking people's reading flow or incurring additional cognitive burden to perform voice-to-author mapping. Importantly, people's reaction to audio cues also depend on their personal preferences and hearing abilities. Some people may want to lower the level of pitch, volume, or duration of audio cues, because they find it disconcerting. Others, however, may prefer to increase pitch, volume, or duration of audio cues and make them more distinctive relative to the screen reader speech so that one does not subsume the other.

Allowing people to customize and personalize the parameters of non-speech audio cues and text-to-speech voices can be a key step towards addressing the issue described above. However, another approach involves rethinking collaborative writing through an activity-centered lens [9, 58] to support the goals a person intends to accomplish and the tasks they are attempting to complete at a particular instance to achieve these goals. For example, are they skimming through the document to understand how other co-authors have contributed? Are they reading to perceive the final state and content of the document? Are they making edits on their own? An individual may not always need continuous awareness of their collaborators' actions, particularly when they are focusing on their own reading or writing activities. Similar to the way visual collaborative interfaces allow users to control the amount of visible collaboration information (e.g., by switching between 'no markup', 'simple markup' and 'all markup' options for tracked changes on Microsoft Word), screen readers could present information relevant to particular tasks (e.g., understanding changes, reading and responding to comments) instead of continuously pushing auditory alerts for collaboration information. One such example may involve having separate "private writing" and "public editing" sessions, as suggested in prior work [62, 89]. Although Wang et al. suggested the separate private and public sessions to support sighted collaborators who want to avoid exposing details of their writing practices [89], here we see that such an activity-centered approach may be helpful for screen reader users to filter out collaboration notifications when they do not need them. Importantly, one's goals and tasks are likely to evolve over time. As such, collaborative tools should determine people's intended tasks either by tracking relevant contextual indicators or by allowing them to declare and switch their current tasks or "modes" fluidly.

9.3 Improving Efficiency in Processing Collaboration Information

One important aspect of efficiency that repeatedly appeared across our study modules is the time required to consume collaboration information. Presenting information sequentially through verbose spoken announcements takes longer to listen to and make sense of the information [67]. In contrast, non-speech audio cues (e.g., earcons, tone overlay) and voice coding can help people quickly process who did what and where by conveying multiple threads of information at once. For

example, the background tone in tone overlay indicates the presence of a comment (or overlapping comments) simultaneously while the commented text is read aloud. Similarly, the voice coding technique reads the content of comments or edits while also highlighting who made that comment or edit. In fact, some participants said that they could get rid of the markup phrases that announce co-authors' names once they learn the corresponding voice mapping in voice coding technique or if they could use personalized voice profiles, thus reducing the time required even further.

While non-speech audio cues and voice coding were decidedly better than the default spoken announcements in terms of time required to understand who edited or commented what and where, the situation gets more complicated when writers need to figure out how the document content has evolved through previous edits. Participants in our study shared that they forget the meaning of a sentence by the time they hear all the spoken announcements for suggested edits and often need to *"go through it a few times"* to piece together how it appears before and after the edits. The contextual presentation technique appeared to reduce cognitive effort in this regard, since participants could more readily perceive how edits altered the meaning of the sentences. This improvement in cognitive effort, however, comes with a compromise in terms of efficiency, as contextual presentation takes longer because it plays a sentence in its original and modified versions. Balancing cognitive overload and efficiency in developing collaboration awareness may require a hierarchical approach with a combination of techniques, where people can opportunistically control what information they will hear in what format depending on their tasks and goals at a particular instance [52]. For example, when someone skims through a document, the presence of edits could be indicated using non-speech audio cues at a higher level (e.g., paragraph level). If the person wants to explore the edits to a specific sentence in more detail, they could use designated keystrokes to listen to the edited text separately or within the context of the sentence.

9.4 Limitations and Future Work

Grounded in findings from an interview study, this paper presents results from a controlled experimental study that investigated the extent to which non-speech audio, voice coding, and contextual presentation support screen reader users' collaboration awareness needs and efficiency relative to the default representations. Our results provide a foundation for future interactive systems that incorporate these techniques and allow for research on other facets of collaborative writing that we were not able to capture within the scope of this controlled study. For instance, with an interactive prototype, future studies may investigate how different representations facilitate (or impede) one's comprehension of collaboration information when they can pause and review, repeat certain comments or edits, and opportunistically query information as they need. Further, a long-term deployment study with an interactive system could evaluate potential learning effects that influence how people perceive collaboration information once they get used to a particular audio representation and use it over time.

Future work could also explore whether non-speech audio cues and voice coding can support visually impaired writers in achieving collaboration awareness and efficiency when they perform real-time editing using screen readers. For instance, our prior work [29] revealed that the lack of awareness around where co-authors are editing in real-time is a key accessibility issue in synchronous editing tools (e.g., Google Docs). Although screen reader users receive spoken notifications when a co-author enters or exits the paragraph they are working on in Google Docs, they do not know the exact location or proximity of the co-author's cursor position relative to their own. As such, screen reader users often avoid close co-editing to reduce the risks of typing over someone else's edits. Furthermore, their own writing gets disrupted by the spoken announcements they hear when a collaborator joins or leaves the document or moves cursor to and from the paragraph

they are working on. Non-speech audio cues may be useful in such cases to provide collaboration information in a less obtrusive manner.

Beyond collaborative writing, our findings are likely to have implications for other collaborative activities such as collaborative programming. Potluri et al. highlighted the challenges with glanceability and alertability in Integrated Development Environments (IDEs), wherein screen reader users cannot process information as easily as a sighted person does with a quick glance at different windows and panes or through real-time visual alerts [67]. Future work could explore whether non-speech audio cues and voice coding techniques can address these issues in the context of collaborative programming.

10 CONCLUSION

With the overarching goal of supporting collaboration within ability-diverse teams, this study set out to rethink the design of collaborative writing tools and address the complexities associated with how screen readers represent collaborative information. Building on the insights we gathered from interviews with 20 visually impaired academics and professionals, we developed auditory representations that indicate collaborative features such as comments and edits in a document using non-speech audio (e.g., earcons and tone overlay), multiple-text-to-speech voices, and contextual presentation techniques. We evaluated these techniques through a controlled study with 48 screen reader users. The results indicate that non-speech audio, voice coding, and contextual presentation perform better than default spoken announcements in conveying complex collaboration information, such as overlapping comments, who commented what, and who edited what in a document. Our analysis also highlights the importance of enabling conditions under which screen reader users can develop collaboration awareness without compromising individual workflows. Moving forward, the use of customizable, context-dependent, and activity-centered representations are most promising for supporting collaboration awareness and efficiency among screen reader users.

ACKNOWLEDGMENTS

This work was supported by NSF grant IIS-1901456 and a gift from Microsoft. We thank our participants for their contributions in the study. We also thank Lou Ann Blake from National Federation of the Blind and Cynthia Bennett for their help with the recruitment process. Additionally, we thank Rawan Mohamed for her help in labeling data from evaluation study sessions and members of the Inclusive Technology Lab and CollabLab for their feedback at various points of this research.

REFERENCES

[1] [n.d.]. Microsoft Custom Voice. https://speech.microsoft.com/customvoice Retrieved May 3, 2020.

[2] [n.d.]. Microsoft Soundscape. https://www.microsoft.com/en-us/research/product/soundscape/ Retrieved May 3, 2020.

[3] [n.d.]. Understanding the Web Content Accessibility Guidelines. https://developer.mozilla.org/en-US/docs/Web/Accessibility/Understanding_WCAG Retrived January 25, 2021.

[4] Ádám Csapó and György Wersényi. 2013. Overview of Auditory Representations in Human-machine Interfaces. *ACM Computing Surveys (CSUR)* 46, 2, Article 19 (December 2013), 23 pages. https://doi.org/10.1145/2543581.2543586

[5] Faisal Ahmed, Yevgen Borodin, Yury Puzis, and I.V. Ramakrishnan. 2012. Why Read if You Can Skim: Towards Enabling Faster Screen Reading. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility* (Lyon, France) *(W4A '12)*. ACM, New York, NY, USA, Article 39, 10 pages. https://doi.org/10.1145/2207016.2207052

[6] Vikas Ashok, Syed Masum Billah, Yevgen Borodin, and IV Ramakrishnan. 2019. Auto-Suggesting Browsing Actions for Personalized Web Screen Reading. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. ACM, New York, NY, USA, 252–260. https://doi.org/10.1145/3320435.3320460

[7] Vikas Ashok, Yury Puzis, Yevgen Borodin, and I.V. Ramakrishnan. 2017. Web Screen Reading Automation Assistance Using Semantic Abstraction. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol,

Cyprus) *(IUI '17)*. ACM, New York, NY, USA, 407–418. https://doi.org/10.1145/3025171.3025229

[8] Ronald M. Baecker, Dimitrios Nastos, Ilona R. Posner, and Kelly L. Mawby. 1993. The User-Centered Iterative Design of Collaborative Writing Software. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. ACM, New York, NY, USA, 399–405. https://doi.org/10.1145/169059.169312

[9] Mark S. Baldwin, Jennifer Mankoff, Bonnie Nardi, and Gillian Hayes. 2020. An Activity Centered Approach to Nonvisual Computer Interaction. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 2, Article 12 (March 2020), 27 pages. https://doi.org/10.1145/3374211

[10] Michael Banf and Volker Blanz. 2013. Sonification of Images for the Visually Impaired Using a Multi-Level Approach. In *Proceedings of the 4th Augmented Human International Conference* (Stuttgart, Germany) *(AH '13)*. ACM, New York, NY, USA, 162–169. https://doi.org/10.1145/2459236.2459264

[11] Jeffrey P. Bigham, Irene Lin, and Saiph Savage. 2017. The Effects of "Not Knowing What You Don't Know" on Web Accessibility for Blind Web Users. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. ACM, New York, NY, USA, 101–109. https://doi.org/10.1145/3132525.3132533

[12] Syed Masum Billah, Vikas Ashok, Donald E. Porter, and I.V. Ramakrishnan. 2017. Ubiquitous Accessibility for People with Visual Impairments: Are We There Yet?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 5862–5868. https://doi.org/10.1145/3025453.3025731

[13] Jeremy Birnholtz and Steven Ibara. 2012. Tracking Changes in Collaborative Writing: Edits, Visibility and Group Maintenance. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. ACM, New York, NY, USA, 809–818. https://doi.org/10.1145/2145204.2145325

[14] Jeremy Birnholtz, Stephanie Steinhardt, and Antonella Pavese. 2013. Write Here, Write Now!: An Experimental Study of Group Maintenance in Collaborative Writing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. ACM, New York, NY, USA, 961–970. https://doi.org/10.1145/2470654.2466123

[15] Tom Boellstorff, Bonnie Nardi, Celia Pearce, and T. L. Taylor. 2013. Words with Friends: Writing Collaboratively Online. *Interactions* 20, 5 (September 2013), 58–61. https://doi.org/10.1145/2501987

[16] Katya Borgos-Rodriguez, Maitraye Das, and Anne Marie Piper. 2021. Melodie: A Design Inquiry into Accessible Crafting through Audio-Enhanced Weaving. *ACM Trans. Access. Comput.* 14, 1, Article 5 (March 2021), 30 pages. https://doi.org/10.1145/3444699

[17] Yevgen Borodin, Jeffrey P. Bigham, Glenn Dausch, and I. V. Ramakrishnan. 2010. More Than Meets the Eye: A Survey of Screen-reader Browsing Strategies. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility* (Raleigh, North Carolina) *(W4A '10)*. ACM, New York, NY, USA, Article 13, 10 pages. https://doi.org/10.1145/1805986.1806005

[18] Yevgen Borodin, Jeffrey P. Bigham, Rohit Raman, and I. V. Ramakrishnan. 2008. What's New? Making Web Page Updates Accessible. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility* (Halifax, Nova Scotia, Canada) *(Assets '08)*. ACM, New York, NY, USA, 145–152. https://doi.org/10.1145/1414471.1414499

[19] Danielle Bragg, Cynthia Bennett, Katharina Reinecke, and Richard Ladner. 2018. A Large Inclusive Study of Human Listening Rates. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 444, 12 pages. https://doi.org/10.1145/3173574.3174018

[20] Stacy M. Branham and Shaun K. Kane. 2015. Collaborative Accessibility: How Blind and Sighted Companions Co-Create Accessible Home Spaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 2373–2382. https://doi.org/10.1145/2702123.2702511

[21] Stacy M. Branham and Shaun K. Kane. 2015. The Invisible Work of Accessibility: How Blind Employees Manage Accessibility in Mixed-Ability Workplaces. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. ACM, New York, NY, USA, 163–171. https://doi.org/10.1145/2700648.2809864

[22] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[23] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

[24] Andy Brown, Caroline Jay, and Simon Harper. 2012. Tailored Presentation of Dynamic Web Content for Audio Browsers. *International Journal of Human-Computer Studies* 70, 3 (2012), 179 – 196. https://doi.org/10.1016/j.ijhcs.2011.11.001

[25] Kenneth P. Burnham and David R. Anderson. 2010. *Model selection and multimodel inference: A practical information-theoretic approach.*

[26] Maria Claudia Buzzi, Marina Buzzi, Barbara Leporini, Giulio Mori, and Victor M. Penichet. 2014. Collaborative Editing: Collaboration, Awareness and Accessibility Issues for the Blind. In *Proceedings of the Confederated International*

Workshops on On the Move to Meaningful Internet Systems: OTM 2014 Workshops - Volume 8842. Springer-Verlag New York, Inc., New York, NY, USA, 567–573. https://doi.org/10.1007/978-3-662-45550-0_58

[27] Michael Connolly, Christof Lutteroth, and Beryl Plimmer. 2010. Document Resizing for Visually Impaired Students. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* (Brisbane, Australia) *(OzCHI '10)*. ACM, New York, NY, USA, 128–135. https://doi.org/10.1145/1952222.1952248

[28] Maitraye Das, Katya Borgos-Rodriguez, and Anne Marie Piper. 2020. Weaving by Touch: A Case Analysis of Accessible Making. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376477

[29] Maitraye Das, Darren Gergle, and Anne Marie Piper. 2019. "It Doesn't Win You Friends": Understanding Accessibility in Collaborative Writing for People with Vision Impairments. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 191 (Nov. 2019), 26 pages. https://doi.org/10.1145/3359293

[30] Paul Dourish and Victoria Bellotti. 1992. Awareness and Coordination in Shared Workspaces. In *Proceedings of the 1992 ACM Conference on Computer-supported Cooperative Work* (Toronto, Ontario, Canada) *(CSCW '92)*. ACM, New York, NY, USA, 107–114. https://doi.org/10.1145/143457.143468

[31] Elizabeth Ellcessor. 2016. *Restricted Access: Media, Disability, and the Politics of Participation*. NYU Press. http://www.jstor.org/stable/j.ctt18040rg

[32] Benoît Encelle, Magali Ollagnier-Beldame, Stéphanie Pouchot, and Yannick Prié. 2011. Annotation-Based Video Enrichment for Blind People: A Pilot Study on the Use of Earcons and Speech Synthesis. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* (Dundee, Scotland, UK) *(ASSETS '11)*. ACM, New York, NY, USA, 123–130. https://doi.org/10.1145/2049536.2049560

[33] Robert S. Fish, Robert E. Kraut, and Mary D. P. Leland. 1988. Quilt: A Collaborative Tool for Cooperative Writing. In *Proceedings of the ACM SIGOIS and IEEECS TC-OA 1988 Conference on Office Information Systems* (Palo Alto, California, USA) *(COCS '88)*. ACM, New York, NY, USA, 30–37. https://doi.org/10.1145/45410.45414

[34] Euan Freeman, Graham Wilson, Stephen Brewster, Gabriel Baud-Bovy, Charlotte Magnusson, and Hector Caltenco. 2017. Audible Beacons and Wearables in Schools: Helping Young Visually Impaired Children Play and Move Independently. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 4146–4157. https://doi.org/10.1145/3025453.3025518

[35] Euan Freeman, Graham Wilson, Dong-Bach Vo, Alex Ng, Ioannis Politis, and Stephen Brewster. 2017. *Multimodal Feedback in HCI: Haptics, Non-Speech Audio, and Their Applications*. ACM and Morgan & Claypool, 277–317. https://doi.org/10.1145/3015783.3015792

[36] Judith Gebauer and Mark Ginsburg. 2009. Exploring the Black Box of Task-Technology Fit. *Commun. ACM* 52, 1 (Jan. 2009), 130–135. https://doi.org/10.1145/1435417.1435447

[37] Andrew Gelman and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

[38] João Guerreiro and Daniel Gonçalves. 2015. Faster Text-to-Speeches: Enhancing Blind People's Information Scanning with Faster Concurrent Speech. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. ACM, New York, NY, USA, 3–11. https://doi.org/10.1145/2700648.2809840

[39] Carl Gutwin and Saul Greenberg. 2002. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work* 11 (2002), 411–446. https://doi.org/10.1023/A:1021271517844

[40] Roy H Hamilton, Alvaro Pascual-Leone, and Gottfried Schlaug. 2004. Absolute Pitch in Blind Musicians. *Neuroreport* 15, 5 (2004), 803–806.

[41] Joe Hutchinson and Oussama Metatla. 2018. An Initial Investigation into Non-Visual Code Structure Overview Through Speech, Non-Speech and Spearcons. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. ACM, New York, NY, USA, Article LBW562, 6 pages. https://doi.org/10.1145/3170427.3188696

[42] Alison Kafer. 2013. *Feminist, Queer, Crip*. Indiana University Press. http://www.jstor.org/stable/j.ctt16gz79x

[43] Grete Helena Kütt, Kevin Lee, Ethan Hardacre, and Alexandra Papoutsaki. 2019. Eye-Write: Gaze Sharing for Collaborative Writing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 497, 12 pages. https://doi.org/10.1145/3290605.3300727

[44] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. https://doi.org/10.2307/2529310

[45] Ida Larsen-Ledet and Henrik Korsgaard. 2019. Territorial Functioning in Collaborative Writing. *Computer Supported Cooperative Work* 28 (2019), 391–433. https://doi.org/10.1007/s10606-019-09359-8

[46] Ida Larsen-Ledet, Henrik Korsgaard, and Susanne Bødker. 2020. Collaborative Writing Across Multiple Artifact Ecologies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376422

[47] Barbara Leporini and Fabio Paterno. 2004. Increasing Usability when Interacting Through Screen Readers. *Universal Access in the Information Society* 3, 1 (March 2004), 57–70. https://doi.org/10.1007/s10209-003-0076-4

[48] Simi Linton. 1998. *Claiming Disability: Knowledge and Identity*. NYU Press. http://www.jstor.org/stable/j.ctt9qfx5w

[49] Paul Benjamin Lowry, Aaron Curtis, and Michelle René Lowry. 2004. Building a Taxonomy and Nomenclature of Collaborative Writing to Improve Interdisciplinary Research and Practice. *The Journal of Business Communication (1973)* 41, 1 (2004), 66–99. https://doi.org/10.1177/0021943603259363

[50] Darren Lunn, Simon Harper, and Sean Bechhofer. 2011. Identifying Behavioral Strategies of Visually Impaired Users to Improve Access to Web Content. *ACM Transactions on Accessible Computing (TACCESS)* 3, 4, Article 13 (April 2011), 35 pages. https://doi.org/10.1145/1952388.1952390

[51] Sergio Mascetti, Andrea Gerino, Cristian Bernareggi, and Lorenzo Picinali. 2017. On the Evaluation of Novel Sonification Techniques for Non-Visual Shape Exploration. *ACM Transactions on Accessible Computing (TACCESS)* 9, 4, Article 13 (April 2017), 28 pages. https://doi.org/10.1145/3046789

[52] Daniel Mendes, Sofia Reis, João Guerreiro, and Hugo Nicolau. 2020. Collaborative Tabletops for Blind People: The Effect of Auditory Design on Workspace Awareness. *Proceedings of the ACM on Human-Computer Interaction* 4, ISS, Article 197 (Nov. 2020), 19 pages. https://doi.org/10.1145/3427325

[53] Oussama Metatla, Sandra Bardot, Clare Cullen, Marcos Serrano, and Christophe Jouffrais. 2020. Robots for Inclusive Play: Co-Designing an Educational Game With Visually Impaired and Sighted Children. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376270

[54] Oussama Metatla, Nick Bryan-Kinns, Tony Stockman, and Fiore Martin. 2012. Cross-modal collaborative interaction between visually-impaired and sighted users in the workplace. In *Proceedings of the 18th International Conference on Auditory Display* (Atlanta, GA, USA) *(ICAD '12)*. 9 pages.

[55] Oussama Metatla, Alison Oldfield, Taimur Ahmed, Antonis Vafeas, and Sunny Miglani. 2019. Voice User Interfaces in Schools: Co-designing for Inclusion with Visually-Impaired and Sighted Pupils. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 378, 15 pages. https://doi.org/10.1145/3290605.3300608

[56] Jonas Moll and Eva-Lotta Sallnäs Pysander. 2013. A Haptic Tool for Group Work on Geometrical Concepts Engaging Blind and Sighted Pupils. *ACM Transactions on Accessible Computing (TACCESS)* 4, 4, Article 14 (July 2013), 37 pages. https://doi.org/10.1145/2493171.2493172

[57] Lourdes Morales, Sonia M. Arteaga, and Sri Kurniawan. 2013. Design Guidelines of a Tool to Help Blind Authors Independently Format Their Word Documents. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. ACM, New York, NY, USA, 31–36. https://doi.org/10.1145/2468356.2468363

[58] Bonnie A. Nardi. 1995. Studying Context: A Comparison of Activity Theory, Situated Action Models, and Distributed Cognition. In *Context and Consciousness: Activity Theory and Human-Computer Interaction*. Massachusetts Institute of Technology, USA, 69–102.

[59] Hugo Nicolau, André Rodrigues, André Santos, Tiago Guerreiro, Kyle Montague, and João Guerreiro. 2019. The Design Space of Nonvisual Word Completion. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. ACM, New York, NY, USA, 249âĂŞ261. https://doi.org/10.1145/3308561.3353786

[60] Uran Oh, Shaun K. Kane, and Leah Findlater. 2013. Follow That Sound: Using Sonification and Corrective Verbal Feedback to Teach Touchscreen Gestures. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (Bellevue, Washington) *(ASSETS '13)*. ACM, New York, NY, USA, Article 13, 8 pages. https://doi.org/10.1145/2513383.2513455

[61] Ricardo Olenewa, Gary M. Olson, Judith S. Olson, and Daniel M. Russell. 2017. Now That We Can Write Simultaneously, How Do We Use That to Our Advantage? *Communication of the ACM* 60, 8 (July 2017), 36–43. https://doi.org/10.1145/2983527

[62] Judith S. Olson, Gary M. Olson, Marianne Storrøsten, and Mark Carter. 1993. Groupwork Close up: A Comparison of the Group Design Process with and without a Simple Group Editor. *ACM Transactions on Information Systems (TOIS)* 11, 4 (Oct. 1993), 321–348. https://doi.org/10.1145/159764.159763

[63] Judith S. Olson, Dakuo Wang, Gary M. Olson, and Jingwen Zhang. 2017. How People Write Together Now: Beginning the Investigation with Advanced Undergraduates in a Project Course. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1, Article 4 (March 2017), 40 pages. https://doi.org/10.1145/3038919

[64] Ignacio Perez-Messina, Claudio Gutierrez, and Eduardo Graells-Garrido. 2018. Organic Visualization of Document Evolution. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. ACM, New York, NY, USA, 497–501. https://doi.org/10.1145/3172944.3173004

[65] Helen Petrie and Omar Kheir. 2007. The Relationship Between Accessibility and Usability of Websites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. ACM, New

York, NY, USA, 397–406. https://doi.org/10.1145/1240624.1240688

[66] Ilona R. Posner and Ronald M. Baecker. 1992. How People Write Together. In *Proceedings of the 25th Hawaii International Conference on System Sciences (HICSS '92)*. 127–138.

[67] Venkatesh Potluri, Priyan Vaithilingam, Suresh Iyengar, Y. Vidya, Manohar Swaminathan, and Gopal Srinivasa. 2018. CodeTalk: Improving Programming Environment Accessibility for Visually Impaired Developers. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3174192

[68] Christopher Power, André Freire, Helen Petrie, and David Swallow. 2012. Guidelines Are Only Half of the Story: Accessibility Problems Encountered by Blind Users on the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. ACM, New York, NY, USA, 433–442. https://doi.org/10.1145/2207676.2207736

[69] Brigitte Röder, Frank Rösler, Erwin Hennighausen, and Fritz Näcker. 1996. Event-Related Potentials During Auditory and Somatosensory Discrimination in Sighted and Blind Human Subjects. *Cognitive Brain Research* 4, 2 (1996), 77–93.

[70] Anna Rouben and Loren Terveen. 2007. Speech and non-speech audio: Navigational information and cognitive load. In *Proceedings of the 13th International Conference on Auditory Display* (Montreal, Canada) *(ICAD '07)*.

[71] Abir Saha and Anne Marie Piper. 2020. Understanding Audio Production Practices of People with Vision Impairments. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 36, 13 pages. https://doi.org/10.1145/3373625.3416993

[72] Daisuke Sato, Shaojian Zhu, Masatomo Kobayashi, Hironobu Takagi, and Chieko Asakawa. 2011. Sasayaki: Augmented Voice Web Browsing Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, New York, NY, USA, 2769–2778. https://doi.org/10.1145/1978942.1979353

[73] John G. Schoeberlein and Yuanqiong Wang. 2014. Usability Evaluation of an Accessible Collaborative Writing Prototype for Blind Users. *Journal of Usability Studies* 10, 1 (Nov. 2014), 26–45. https://dl.acm.org/doi/10.5555/2817310.2817313

[74] Brian A. Smith and Shree K. Nayar. 2018. The RAD: Making Racing Games Equivalently Accessible to People Who Are Blind. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 516, 12 pages. https://doi.org/10.1145/3173574.3174090

[75] Amanda Stent, Ann Syrdal, and Taniya Mishra. 2011. On the Intelligibility of Fast Synthesized Speech for Individuals with Early-Onset Blindness. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* (Dundee, Scotland, UK) *(ASSETS '11)*. ACM, New York, NY, USA, 211–218. https://doi.org/10.1145/2049536.2049574

[76] Kevin M. Storer and Stacy M. Branham. 2019. "That's the Way Sighted People Do It": What Blind Parents Can Teach Technology Designers About Co-Reading with Children. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. ACM, New York, NY, USA, 385–398. https://doi.org/10.1145/3322276.3322374

[77] James Tam and Saul Greenberg. 2006. A Framework for Asynchronous Change Awareness in Collaborative Documents and Workspaces. *International Journal of Human-Computer Studies* 64, 7 (July 2006), 583–598. https://doi.org/10.1016/j.ijhcs.2006.02.004

[78] Mary Frances Theofanos and Janice (Ginny) Redish. 2003. Bridging the Gap: Between Accessibility and Usability. *Interactions* 10, 6 (November 2003), 36–51. https://doi.org/10.1145/947226.947227

[79] Anja Thieme, Cecily Morrison, Nicolas Villar, Martin Grayson, and Siân Lindley. 2017. Enabling Collaboration in Learning Computer Programing Inclusive of Children with Vision Impairments. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. ACM, New York, NY, USA, 739–752. https://doi.org/10.1145/3064663.3064689

[80] Brianna J. Tomlinson, Jared Batterman, Yee Chieh Chew, Ashley Henry, and Bruce N. Walker. 2016. Exploring Auditory Graphing Software in the Classroom: The Effect of Auditory Graphs on the Classroom Environment. *ACM Transactions on Accessible Computing (TACCESS)* 9, 1, Article 3 (Nov. 2016), 27 pages. https://doi.org/10.1145/2994606

[81] Shannon M. Tomlinson. 2016. Perceptions of Accessibility and Usability by Blind or Visually Impaired Persons: A Pilot Study. In *Proceedings of the 79th ASIST Annual Meeting: Creating Knowledge, Enhancing Lives Through Information & Technology* (Copenhagen, Denmark) *(ASIST '16)*. American Society for Information Science, Silver Springs, MD, USA, Article 120, 4 pages.

[82] Johnny Torres, Sixto García, and Enrique Peláez. 2019. Visualizing Authorship and Contribution of Collaborative Writing in E-Learning Environments. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. ACM, New York, NY, USA, 324–328. https://doi.org/10.1145/3301275.3302328

[83] Selen Türkay, Daniel Seaton, and Andrew M. Ang. 2018. Itero: A Revision History Analytics Tool for Exploring Writing Behavior and Reflection. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. ACM, New York, NY, USA, Article LBW052, 6 pages. https://doi.org/10.1145/3170427.3188474

[84] Markel Vigo and Simon Harper. 2013. Challenging Information Foraging Theory: Screen Reader Users Are Not Always Driven by Information Scent. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media* (Paris, France) *(HT '13)*. ACM, New York, NY, USA, 60–68. https://doi.org/10.1145/2481492.2481499

[85] Markel Vigo and Simon Harper. 2013. Coping Tactics Employed by Visually Disabled Users on the Web. *International Journal of Human-Computer Studies* 71, 11 (November 2013), 1013–1025. https://doi.org/10.1016/j.ijhcs.2013.08.002

[86] Patrice Voss, Maryse Lassonde, Frederic Gougoux, Madeleine Fortin, Jean-Paul Guillemot, and Franco Lepore. 2004. Early- and Late-Onset Blind Individuals Show Supra-Normal Auditory Abilities in Far-Space. *Current Biology* 14, 19 (2004), 1734–1738.

[87] Herman Wahidin, Jenny Waycott, and Steven Baker. 2018. The Challenges in Adopting Assistive Technologies in the Workplace for People with Visual Impairments. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction* (Melbourne, Australia) *(OzCHI '18)*. ACM, New York, NY, USA, 432–442. https://doi.org/10.1145/3292147.3292175

[88] Dakuo Wang, Judith S. Olson, Jingwen Zhang, Trung Nguyen, and Gary M. Olson. 2015. DocuViz: Visualizing Collaborative Writing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 1865–1874. https://doi.org/10.1145/2702123.2702517

[89] Dakuo Wang, Haodan Tan, and Tun Lu. 2017. Why Users Do Not Want to Write Together When They Are Writing Together: Users' Rationales for Today's Collaborative Writing Practices. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW, Article 107 (Dec. 2017), 18 pages. https://doi.org/10.1145/3134742

[90] Susan Wendell. 1996. *The Rejected Body: Feminist Philosophical Reflections on Disability*. Routledge.

[91] Michele A. Williams, Caroline Galbraith, Shaun K. Kane, and Amy Hurst. 2014. "Just Let the Cane Hit It": How the Blind and Sighted See Navigation Differently. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility* (Rochester, New York, USA) *(ASSETS '14)*. ACM, New York, NY, USA, 217–224. https://doi.org/10.1145/2661334.2661380

[92] Graham Wilson and Stephen A. Brewster. 2016. Using Dynamic Audio Feedback to Support Peripersonal Reaching in Young Visually Impaired People. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* (Reno, Nevada, USA) *(ASSETS '16)*. ACM, New York, NY, USA, 209–218. https://doi.org/10.1145/2982142.2982160

[93] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. 2007. The Increasing Dominance of Teams in Production of Knowledge. *Science* 316, 5827 (2007), 1036–1039. https://doi.org/10.1126/science.1136099

[94] Soobin Yim, Dakuo Wang, Judith Olson, Viet Vu, and Mark Warschauer. 2017. Synchronous Collaborative Writing in the Classroom: Undergraduates' Collaboration Practices and Their Impact on Writing Style, Quality, and Quantity. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 468–479. https://doi.org/10.1145/2998181.2998356

[95] Chien Wen Yuan, Benjamin V. Hanrahan, Sooyeon Lee, Mary Beth Rosson, and John M. Carroll. 2017. "I Didn't Know That You Knew I Knew": Collaborative Shopping Practices Between People with Visual Impairment and People with Vision. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW, Article 118 (December 2017), 18 pages. https://doi.org/10.1145/3134753

[96] Dongsong Zhang, Lina Zhou, Judith O. Uchidiuno, and Isil Y. Kilic. 2017. Personalized Assistive Web for Improving Mobile Web Browsing and Accessibility for Visually Impaired Users. *ACM Transactions on Accessible Computing (TACCESS)* 10, 2, Article 6 (April 2017), 22 pages. https://doi.org/10.1145/3053733

[97] Yeshuang Zhu, Shichao Yue, Chun Yu, and Yuanchun Shi. 2017. CEPT: Collaborative Editing Tool for Non-Native Authors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 273–285. https://doi.org/10.1145/2998181.2998306

A DETAILS OF INTERVIEW PARTICIPANTS

Table 2. **Participant information (all names are pseudonyms)**

Name	Self-reported Visual Ability	Occupation	Documents produced
Addison	Totally blind since birth	Customer service assistant, assistive tech instructor, blogger	Assistive tech manuals, tutorials, books
Alex	Legally blind from Retinitis Pigmentosa, gradual vision loss	PhD student, accessibility researcher	Research papers

Table 2. **Participant information (all names are pseudonyms)**

Name	Self-reported Visual Ability	Occupation	Documents produced
Bella	Nearly totally blind since birth, some light perception in one eye	Assistive tech trainer, blogger	Website content, presentations, papers
Bill	Profound vision impairment, some light perception in one eye, gradual vision loss	Entrepreneur, accessibility consultant	Research papers, website content, blog posts, books
Daniel	Totally blind since birth due to glaucoma	Accessibility consultant, blogger, (past: customer tech support)	Technical articles, assistive tech related articles
David	Nearly totally blind since birth, some light perception	Contract employee (quality assurance, usability testing)	Assistive tech articles, (past: course projects)
Elena	Nearly totally blind since birth, some light perception	Accessibility and assistive tech specialist	Assistive tech related grant proposals
Emma	Legally blind, nearly functional print vision in one eye, born with cataract, developed glaucoma	Accessibility and assistive tech specialist	Assistive tech related articles
Ethan	Totally blind since 12 years old	Business trading analyst, blogger	Business report, technical guides, website content
Grace	Totally blind since 19 years old	Digital accessibility consultant	Meeting notes, project proposals, assistive tech related articles
Henry	Nearly totally blind since 9 years old, some light perception	Accessibility consultant, blogger, entrepreneur	Event planning documents
Isaac	Nearly totally blind since birth, light perception in one eye	PhD student, accessibility researcher, (past: research intern)	Research papers, course projects, reports
Kaylee	Totally blind since birth	Applied Sciences Degree student	Exam papers, shopping list
Lily	Nearly totally blind since birth due to Retinopathy of prematurity, light perception in one eye	BS student	Course projects
Maya	Totally blind for 12 years	PhD student, accessibility researcher, activist	Research papers, class projects, social events
Mila	Totally blind since birth due to retinopathy of prematurity	Museum consultant, researcher	Research papers, books
Nathan	Legally blind from Retinitis Pigmentosa, gradual vision loss	Research assistant	Research papers, dissertation, course projects
Nova	Nearly totally blind, some light perception, born with retinopathy of prematurity and glaucoma, gradual vision loss	Attorney, accessibility advocate, assistive tech analyst	Legal documents
Ryan	Nearly totally blind since birth, some light perception	Grad student, (past: intern at law firm)	Course projects, court orders
Sofia	Legally blind due to congenital glaucoma, some light perception, gradual vision loss	Customer tech Support (work from home)	Help center documentation, assistive tech user guides

B STATISTICAL ANALYSES RESULTS

Table 3. **Module 1: Results of linear mixed effects logistic regression on performance measures. Reference technique: default announcement, reference order level: 1, complex: high complexity, usage: frequent comments usage.**

Question	Model Info	Predictor	*log(OR)*	SE	*p*
Where are the comments distributed?	AIC:201.5, df.resid:278	Intercept	2.92	0.72	<0.001***
		Earcons	-1.10	0.72	0.12
		Tone overlay	-0.56	0.77	0.47
		Complex	-1.40	0.70	0.046*
		Order:2	0.58	0.49	0.23
		Order:3	0.32	0.46	0.49
		Usage	-0.74	0.44	0.09†
		Earcons x Complex	**2.95**	1.08	0.006**
		Tone overlay x Complex	**1.96**	1.04	0.059†
Are there any overlapping comments?	AIC:190.5, df.resid:280	Intercept	2.03	0.58	<0.001***
		Earcons	0.33	0.47	0.49
		Tone overlay	0.77	0.53	0.15
		Complex	-0.08	0.41	0.84
		Order:2	-0.32	0.47	0.49
		Order:3	0.45	0.55	0.42
		Usage	0.06	0.44	0.89
Comprehension of the passage content	AIC:383.7, df.resid:280	Intercept	0.70	0.39	0.07†
		Earcons	-0.38	0.31	0.22
		Tone overlay	**0.63**	0.33	0.055†
		Complex	-0.07	0.26	0.80
		Order:2	-0.85	0.32	0.008**
		Order:3	-0.17	0.32	0.61
		Usage	0.08	0.33	0.80

***p<0.001, **p<0.01, *p<0.05, and †p<0.10

Table 4. **Module 1: Results of linear mixed effects regression on statements with 5-point Likert ratings and overall preference on a scale of 1-3. Reference technique: default announcement, reference order level: 1, usage: frequent comments usage.**

Statement	Model Info	Predictor	β	SE	p
I could easily understand where the comments were attached.	AIC:361.9, df.resid:136	Intercept	4.38	0.18	<0.001***
		Earcons	-0.02	0.15	0.89
		Tone overlay	-1.3e-15	0.15	1.00
		Order:2	-0.06	0.15	0.68
		Order:3	-0.15	0.15	0.33
		Usage	0.22	0.17	0.20
I could easily understand if there were any overlapping comments.	AIC:378.7, df.resid:136	Intercept	3.96	0.19	<0.001***
		Earcons	0.27	0.15	0.08†
		Tone overlay	**0.58**	0.15	<0.001***
		Order:2	-0.25	0.15	0.11
		Order:3	-0.27	0.15	0.08†
		Usage	0.22	0.18	0.24
This technique was easy to learn.	AIC:377.1, df.resid:136	Intercept	4.05	0.18	<0.001***
		Earcons	0.10	0.17	0.54
		Tone overlay	-2.9e-15	0.17	1.00
		Order:2	-0.08	0.17	0.62
		Order:3	-2.9e-15	0.17	1.00
		Usage	0.14	0.15	0.36
Understanding this technique required a lot of mental effort.	AIC:471.0, df.resid:136	Intercept	2.53	0.26	<0.001***
		Earcons	0.17	0.22	0.44
		Tone overlay	-0.27	0.22	0.21
		Order:2	0.25	0.22	0.25
		Order:3	0.27	0.22	0.21
		Usage	0.08	0.25	0.75
This technique disrupted my reading flow.	AIC:479.7, df.resid:136	Intercept	3.71	0.26	<0.001***
		Earcons	**-0.50**	0.24	0.04*
		Tone overlay	**-1.02**	0.24	<0.001***
		Order:2	0.13	0.24	0.60
		Order:3	0.10	0.24	0.66
		Usage	0.11	0.23	0.63
Overall preference (higher is better)	AIC:346.2, df.resid:138	Intercept	1.72	0.14	<0.001***
		Earcons	0.22	0.16	0.17
		Tone overlay	**0.63**	0.16	<0.001***
		Usage	0.00	0.13	1.00

***p<0.001, **p<0.01, *p<0.05, and †p<0.10

Table 5. **Module 2: Results of linear mixed effects logistic regression on performance measures. Reference technique: default consistent voice, reference order level: 1, complex: high complexity, usage: frequent comments usage.**

Question	Model Info	Predictor	*log(OR)*	SE	*p*
Who commented about a specific text?	AIC:233.1, df.resid:186	Intercept	-0.54	0.41	0.20
		Voice coding	**1.43**	0.36	<0.001***
		Complex	-1.62	0.37	<0.001***
		Order:2	-0.02	0.34	0.96
		Usage	0.23	0.40	0.57
What did the commenter say about the specific text?	AIC:263.0, df.resid:186	Intercept	-0.56	0.41	0.18
		Voice coding	**1.03**	0.33	0.002**
		Complex	-0.25	0.32	0.43
		Order:2	0.35	0.32	0.27
		Usage	-0.21	0.39	0.60
Who commented the most?	AIC:214.0, df.resid:185	Intercept	2.17	0.63	<0.001***
		Voice coding	0.58	0.60	0.33
		Complex	-0.57	0.54	0.29
		Order:2	-1.37	0.41	<0.001***
		Usage	0.29	0.51	0.57
		Voice coding x Complex	**-1.77**	0.79	0.02*
Are there any replies to a specific comment?	AIC:218.8, df.resid:186	Intercept	0.49	0.41	0.23
		Voice coding	0.30	0.35	0.39
		Complex	1.13	0.37	0.002**
		Order:2	0.06	0.35	0.86
		Usage	-0.05	0.38	0.89
Comprehension of the passage content	AIC:201.3, df.resid:187; does not include *order* for model convergence	Intercept	1.91	0.46	<0.001***
		Voice coding	**-0.81**	0.38	0.03*
		Complex	-0.13	0.37	0.71
		Usage	0.02	0.40	0.95

***p<0.001, **p<0.01, *p<0.05, and $^{\dagger}p$<0.10

Table 6. **Module 2: Results of linear mixed effects regression on statements with 5-point Likert ratings and overall preference on a scale of 1-2. Reference technique: default consistent voice, reference order level: 1, usage: frequent comments usage.**

Statement	Model Info	Predictor	β	SE	p
I could easily understand who commented what.	AIC:265.4, df.resid:90	Intercept	4.09	0.21	<0.001***
		Voice coding	0.17	0.14	0.26
		Order:2	-0.08	0.14	0.57
		Usage	0.20	0.24	0.42
I could easily understand what the comment was about.	AIC:199.8, df.resid:90	Intercept	4.19	0.17	<0.001***
		Voice coding	0.15	0.09	0.10
		Order:2	-0.10	0.09	0.24
		Usage	0.32	0.20	0.12
I could easily understand the replies to a comment.	AIC:207.6, df.resid:90	Intercept	4.07	0.19	<0.001***
		Voice coding	**0.19**	0.08	0.02*
		Order:2	-0.23	0.08	0.006**
		Usage	0.45	0.24	0.07†
This technique was easy to learn.	AIC:217.0, df.resid:90	Intercept	4.26	0.17	<0.001***
		Voice coding	0.21	0.11	0.07†
		Order:2	-0.21	0.11	0.07†
		Usage	-0.06	0.19	0.76
Understanding this technique required a lot of mental effort.	AIC:315.1, df.resid:90	Intercept	2.67	0.29	<0.001***
		Voice coding	-0.21	0.17	0.24
		Order:2	0.13	0.17	0.47
		Usage	0.27	0.34	0.44
This technique disrupted my reading flow.	AIC:309.5, df.resid:90	Intercept	3.11	0.26	<0.001***
		Voice coding	**-0.42**	0.19	0.03*
		Order:2	0.25	0.19	0.20
		Usage	0.30	0.29	0.30
Overall preference (higher is better)	AIC:102.6, df.resid:91	Intercept	1.20	0.08	<0.001***
		Voice coding	**0.60**	0.08	<0.001***
		Usage	-4.3e-16	0.08	1.0

***p<0.001, **p<0.01, *p<0.05, and $^{\dagger}p$<0.10

Table 7. **Module 3: Results of linear mixed effects logistic regression on performance measures. Reference technique: default announcement, reference order level: 1, complex: high complexity, usage: frequent edits usage.**

Question	Model Info	Predictor	*log(OR)*	SE	*p*
Who edited a specific sentence?	AIC:285.1, df.resid:280	Intercept	-0.87	0.47	$0.07^{\dagger}$
		Contextual	**1.12**	0.42	0.007^{**}
		Contextual voice coding	**1.48**	0.42	$<0.001^{***}$
		Complex	-2.84	0.41	$<0.001^{***}$
		Order:2	0.64	0.40	0.11
		Order:3	0.34	0.40	0.40
		Usage	-0.38	0.42	0.36
How did the meaning of the sentence alter after the edits?	AIC:296.7, df.resid:280	Intercept	-1.33	0.47	0.005^{**}
		Contextual	**1.73**	0.43	$<0.001^{***}$
		Contextual voice coding	**2.73**	0.47	$<0.001^{***}$
		Complex	-2.44	0.37	$<0.001^{***}$
		Order:2	0.95	0.39	0.02^{*}
		Order:3	0.29	0.39	0.46
		Usage	-0.33	0.40	0.40
Who edited the most?	AIC:360.5, df.resid:283; does not include *order* and *usage* for model convergence	Intercept	0.98	0.37	0.008^{**}
		Contextual	-0.31	0.35	0.38
		Contextual voice coding	-0.12	0.35	0.72
		Complex	-1.16	0.29	$<0.001^{***}$

$^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$, and $^{\dagger}p<0.10$

Table 8. **Module 3: Results of linear mixed effects regression on statements with 5-point Likert ratings and overall preference on a scale of 1-3. Reference technique: default announcement, reference order level: 1, usage: frequent edits usage.**

Statement	Model Info	Predictor	β	SE	p
I could easily understand who edited what.	AIC:426.4, df.resid:136	Intercept	3.65	0.22	<0.001***
		Contextual	0.19	0.17	0.26
		Contextual voice coding	**0.38**	0.17	0.03*
		Order:2	0.06	0.17	0.71
		Order:3	6.3e-15	0.17	1.00
		Usage	-0.09	0.26	0.74
I could easily understand how edits altered the meaning of a sentence	AIC:437.0, df.resid:136	Intercept	3.31	0.22	<0.001***
		Contextual	0.23	0.18	0.22
		Contextual voice coding	**0.58**	0.18	0.002**
		Order:2	0.10	0.18	0.57
		Order:3	0.08	0.18	0.65
		Usage	-0.38	0.23	0.11
This techniques was easy to learn.	AIC:404.6, df.resid:136	Intercept	3.47	0.19	<0.001***
		Contextual	0.23	0.17	0.18
		Contextual voice coding	**0.42**	0.17	0.02*
		Order:2	0.08	0.17	0.62
		Order:3	0.06	0.17	0.71
		Usage	-0.30	0.20	0.14
Understanding this technique required a lot of mental effort.	AIC:431.8, df.resid:136	Intercept	3.74	0.22	<0.001***
		Contextual	-0.31	0.18	0.08†
		Contextual voice coding	**-0.73**	0.18	<0.001***
		Order:2	0.21	0.18	0.24
		Order:3	0.06	0.18	0.72
		Usage	0.19	0.24	0.44
This technique disrupted my reading flow	AIC:409.8, df.resid:136	Intercept	4.05	0.20	<0.001***
		Contextual	-0.29	0.18	0.10
		Contextual voice coding	**-0.83**	0.18	<0.001***
		Order:2	-0.06	0.18	0.72
		Order:3	-0.25	0.18	0.16
		Usage	0.63	0.19	0.002**
Overall preference (higher is better)	AIC:341.2, df.resid:138	Intercept	1.66	0.13	<0.001***
		Contextual	**0.32**	0.15	0.04*
		Contextual voice coding	**0.71**	0.15	<0.001***
		Usage	0.00	0.13	1.00

***p<0.001, **p<0.01, *p<0.05, and †p<0.10