VoxVista: Enhancing Screen Reading Experience for Online User Comments

Yash Prakash
Old Dominion University
Department of Computer Science
Norfolk, Virginia, USA
yprak001@odu.edu

Akshay Kolgar Nayak
Old Dominion University
Department of Computer Science
Norfolk, Virginia, USA
anaya001@odu.edu

Shoaib Alyaan
Old Dominion University
Department of Computer Science
Norfolk, Virginia, USA
salya006@odu.edu

Sampath Jayarathna
Old Dominion University
Department of Computer Science
Norfolk, Virginia, USA
ujayarat@odu.edu

Hae-Na Lee
Michigan State University
Department of Computer Science
East Lansing, Michigan, USA
leehaena@msu.edu

Vikas Ashok
Old Dominion University
Department of Computer Science
Norfolk, Virginia, USA
vganjigu@odu.edu

Abstract

Online discussions have become integral to how people exchange ideas, form opinions, and participate in collective deliberation. While sighted users can comfortably engage with online discussions, blind users who are dependent on screen readers are forced to listen to long threads narrated in a single, monotonic voice that lacks prosodic variation, rhythm, or emotion. This robotic auditory experience not only deteriorates the user engagement with the content but also increases cognitive strain, by making it difficult to remain attentive and discern meaning beyond literal words. In an interview study, most blind participants reported that monotonous narration hindered their ability to detect salient information, perceive emotional cues, and comprehend content authors' intents in discussions. Many described experiencing mental fatigue when listening to 'flat', 'uninspiring' voices, noting that their attention tended to diminish quickly over time. The participants also indicated that they often tried to 'add' prosodic variation or emotional inflection themselves in their minds, but characterized this compensatory effort as mentally taxing and cognitively demanding. To address this issue, we introduce VoxVista, a multi-voice design framework driven by a large language model that leverages a custom voice-preference dataset to assign personalized voice profiles to user posts in discussions, thereby replacing the traditional monotone narration in screen readers with a more expressive, dynamic, and contextually-aware narration. In a study with 20 blind participants, we observed that VoxVista significantly improved user engagement, comprehension, and willingness to continue listening to longer discussions.

CCS Concepts

• **Human-centered computing** → **Accessibility technologies**; **Auditory feedback**.

CHIIR '26, Seattle, WA, USA

ACM ISBN 979-8-4007-2414-5/2026/03
https://doi.org/10.1145/3786304.3788850

Keywords

User experience, User generated comments, Blind, Visual impairment, Screen reader

ACM Reference Format:
Yash Prakash, Akshay Kolgar Nayak, Shoaib Alyaan, Sampath Jayarathna, Hae-Na Lee, and Vikas Ashok. 2026. VoxVista: Enhancing Screen Reading Experience for Online User Comments. In *2026 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '26), March 22–26, 2026, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3786304.3788850

1 Introduction

Digitization and socio-technological innovations have enabled online users to actively create and share content, positioning user-generated content, particularly user discussions, as a key area of focus in communication studies [22, 40]. Online forums provide a platform for users to share experiences, exchange opinions, and foster community-building through organic conversations [38]. People generally read comments for a variety of purposes: to understand and compare perspectives, for entertainment, to gain updates on a topic, or to gauge the general community sentiment [30].

For sighted users, visually skimming comments is a fast and intuitive process, guided by visual cues like indentation, font size, and colors. These cues help them quickly grasp conversation hierarchy (i.e., parent-child relationships), understand a conversation's flow, and detect salient points. Sighted users can also gauge emotional tone through emojis and punctuation, allowing them to assess sentiment at a glance. However, blind users perceive user comments through a different lens, one shaped by sound rather than sight [26] (see Figure 1). Relying on screen readers like JAWS or VoiceOver, blind users listen to text after it is transformed into synthesized speech, often delivered in a single monotonic voice, with limited expression [11] and diversity [10]. This monotony makes it difficult to discern the flow of conversation. The lack of variations in generated voice turns dialog into a blur of words, leading to confusion and increased cognitive load as users struggle to mentally untangle conversations that should otherwise be processed with ease.

To uncover the auditory usability challenges faced by blind users in online discussions, we conducted an interview study with 14

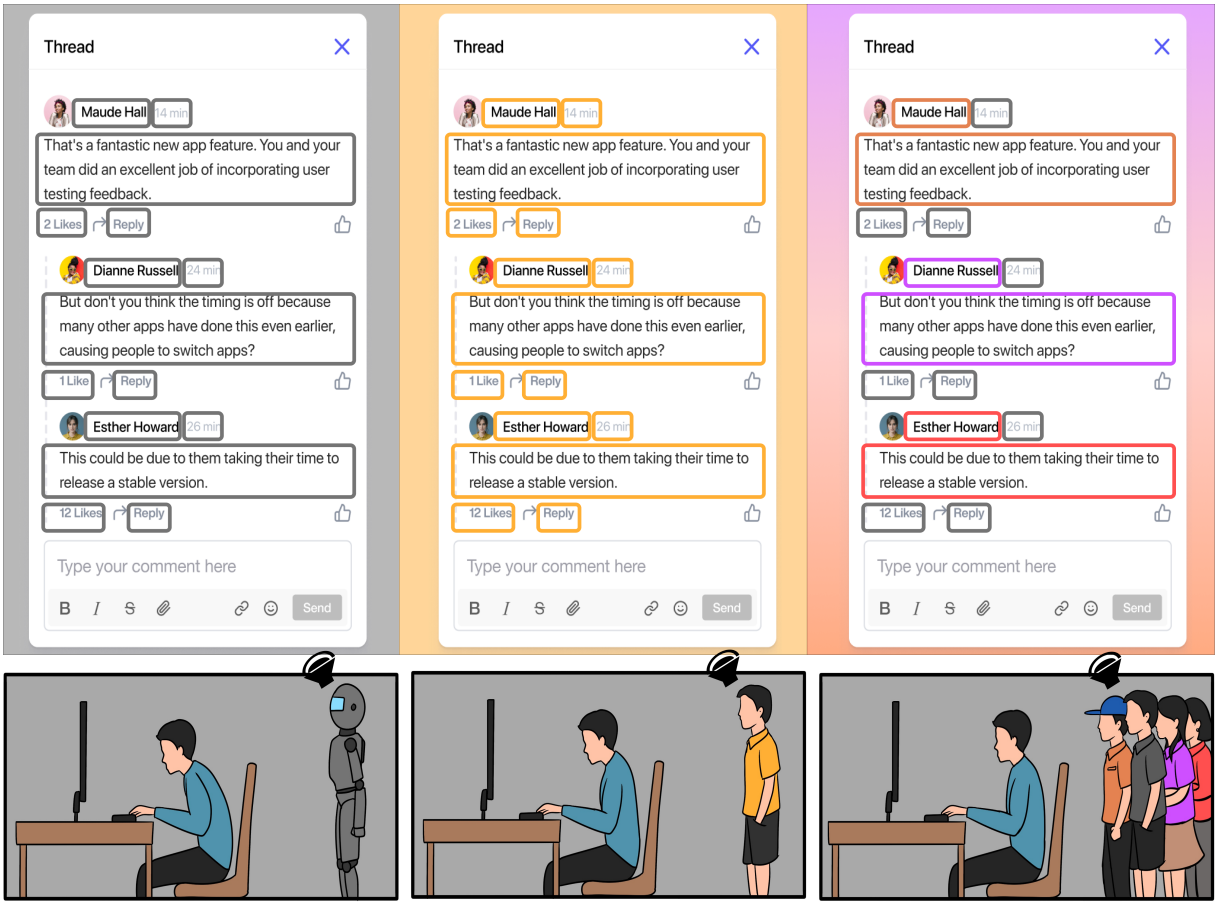

Figure 1: (Left & middle) Prior research has focused on evaluating user experiences with single robotic or natural voices, particularly in text-to-speech systems (screen readers) used by visually impaired individuals [26]. (Right) Our study expands this research by exploring the potential of multi-voice systems to enhance the overall experience of blind users when interacting with online comment sections, offering a more dynamic and contextually rich auditory experience.

blind participants. All participants reported that navigating discussion threads was mentally exhausting due to three main factors: (1) monotonous narration; (2) the need to self-apply inflections or variations internally; and (3) the need to mentally segment continuous speech for tracking the conversation flow.

To improve the screen reading experience for user comments, we introduce VoxVista, a design framework that leverages a large language model-based agent acting as a user within the comment section. The agent's task is to understand how a given comment is entangled with other comments in the dialog, identify the conversational flow, and then assign an appropriate voice profile to each comment based on contextual understanding. At a high level, VoxVista replaces the conventional monotone narration of screen readers with a more expressive, dynamic, and context-aware narration. Moreover, VoxVista employs distinct voices (Figure 1) to represent different commenters within a discussion thread, enabling blind users to more easily follow the conversational flow and identify speaker context.

We conducted a Voice Experience Comparative Study with 20 participants to compare different voice configurations – single robotic voice, single natural voice, and multi-voice with natural tones. The results clearly favored the multi-voice configuration; all participants reported that this configuration significantly enhanced their experience, by enabling them to perceive comments in a more engaging manner. In sum, our contributions are: (i) A custom voice-preference dataset capturing the needs of users regarding online comments; and (ii) Design and evaluation of VoxVista framework that dynamically assigns profile-based voices to user comments.

2 Related Work

2.1 Voice-based Digital User Experience

Prior research in HCI, speech technology, and communication have shown that the qualities and characteristics of a voice, whether human recorded or synthetic text-to-speech (TTS), can influence the dynamics of interaction [9, 10, 23, 33]. While TTS is often an optional modality for sighted users when consuming digital content,

blind and visually impaired (BVI) users rely heavily on TTS for most or all of their digital interactions. Prior research exploring the use of TTS in screen readers has predominantly focused on optimizing listening speed [4, 5, 7, 8], utilizing concurrent speech [14, 15, 20] to improve the efficiency of information consumption, and generating more natural-sounding voices [13, 39].

However, research pertaining to TTS requirements and voice preferences for BVI users is relatively scarce [26]. Podsiadlo et al. [26] conducted a study to explore the challenges of BVI users regarding TTS technology, and found that during leisure activities such as reading books or news articles, the intelligibility of the voice was the primary factor influencing user experience. Participants noted that while the acoustic qualities of the voice and its gender were less important, it was crucial for the voice to be pleasant, particularly for extended listening sessions. Additionally, participants expressed a preference for neutral emotional expression in the voice, allowing them to project their own emotions onto the content material, akin to the process of silent reading. However, Podsiadlo et al.'s [26] findings were based on the use of a single voice, leaving two critical research gaps. First, there is limited understanding of how blind users perceive and benefit from multiple distinct voices in a single listening session. Second, the scope of earlier works did not specifically consider discussion forums and comment sections, which have distinct characteristics compared to traditional reading materials. Unlike books or news articles where content is authored by one or few voices and follows a cohesive narrative, online discussions involve multiple participants with diverse perspectives, conversational dynamics, and varying emotional tones. These multi-party exchanges require listeners to track who said what, understand conversational flow, and differentiate between speakers—tasks that may be better supported by multi-voice rendering rather than single-voice narration. To address these gaps, we conducted a study assessing our VoxVista prototype with 20 participants, comparing preferences for default robotic voice, a single natural voice, and dynamic multi-voice profiles specifically in the context of user-generated comments and discussion forums.

2.2 Screen Reader Interaction in Online Discussion Forums

Prior research on online user-generated comments has extensively explored areas such as the influence of comments on readers' perceptions of the associated article [36, 40], behavioral engagement through emotional responses [29, 44], persuasiveness of comments [16, 41], and the overall impact of comments on public opinion [19, 37]. However, studies aimed at enhancing the user experience of comment sections are scarce, with most research focusing predominantly on sighted users [1, 38]. Research addressing the experiences of BVI users in this context remains limited [12, 31, 32, 34, 35]. Sunkara et al. [32] in their interview study found participants found it tedious and frustrating to navigate discussion forums due to redundancy, high volume, and limited customization. Repeated posts, acknowledgments, and off-topic comments increased listening time and distracted from the main discussion. Unlike sighted users who can skim text visually, blind users face greater challenges. To address this, researchers designed a chrome extension that allowed blind users to filter posts based on personal preferences.

While this is an important step in improving screen reader user experience in comment sections, another aspect of UX (i.e., auditory perception) remains under-explored, despite its potential to significantly enhance user enjoyment, comprehension, and the willingness to continue listening. To address this, we introduce VoxVista, enabling screen reader users to experience comments as if they are listening to humans engaging in a conversation, thereby enhancing the overall interaction with dynamic auditory feedback.

3 Blind Usability Issues with Discussion Forums

We conducted an IRB-approved interview study with 14 blind screen reader users[1] (5 female, 9 male) to investigate their interaction challenges with online discussion forums. All participants were experienced TTS users with diverse educational backgrounds and regular engagement with online forums. The study focused on understanding participants' experiences, including how often they use forums, the contexts in which they interact with them, and the platforms they most frequently engage with. The zoom interviews lasted 45–75 minutes, and were audio-recorded and transcribed. Based on initial exploration and prior work, we designed seed questions to probe both general usability and auditory-specific challenges:

- How does the style or tone of auditory narration affect your comprehension of comments?
- Do you experience fatigue or difficulty focusing when listening to comments for extended periods?
- How do you handle emotionally neutral or monotonous speech in comments?
- Do you consciously adapt or "imagine" speech while listening to understand content better?

The transcribed interviews were analyzed using an open coding approach [28], where we iteratively examined participants' responses to identify recurring patterns, usability obstacles, and adaptive strategies. This method allowed us to capture both explicit challenges reported by participants and implicit behaviors adopted while navigating forums with a screen reader.

Effects of Monotonous Narration on Comprehension and Engagement. Most participants (12) reported that monotonous narration in auditory descriptions of user-generated comments impaired their ability to identify key information. Several participants noted losing focus during prolonged listening, with some stating they had to increase the pace of listening due to boredom, and others indicating that their interest waned over time. When auditory narration lacks variation in tone, rhythm, and emphasis, listeners miss essential cues that normally help highlight critical content, segment information, and convey the author's intent. Consequently, comprehension relies more heavily on controlled cognitive processes such as verbal working memory and attention-based monitoring [25]. In addition, participants (3) mentioned difficulty interpreting emotional cues and understanding the intentions of comment authors when the speech was monotonous. Expressive narration, by contrast, can sustain attention, enhance understanding, and improve retention of nuanced meaning and sentiment [24].

[1]Participant demographics are available at: https://github.com/accessodu/VoxVista.git

Monotony-Induced Cognitive Fatigue and Effortful Mental Strategies. Many participants (10) described experiencing mental fatigue when listening to flat or uninspiring voices. The absence of variation in tone, pitch, or emphasis increases listening effort and diminishes comprehension. For blind users, who rely exclusively on auditory information, this effect is amplified, as expressive speech can often compensate for the lack of visual context [18]. A few participants (5) also reported actively attempting to self-apply variation or emotional inflection into their inner speech to better understand comments. While this strategy helped comprehension, it was mentally taxing. Such deliberate mental modulation indicates that listeners are engaging higher-order cognitive mechanisms to supplement missing auditory cues, highlighting the extra effort required to process monotonous speech [3].

Mentally segmenting monotonous speech to follow content. Several participants (7) reported that when user comments were read with a single monotonous voice, they needed to consciously segment the auditory stream into smaller units to understand the content. Without natural variations in tone or emphasis, it was difficult to identify boundaries between comments or detect which points were important. Four participants explained that they often had to mentally 'pause' after each comment and summarize it internally before moving on, effectively creating their own structure to keep track of the discussion. Two participants noted that this effort often lead to fatigue in longer threads.

4 VoxVista Design Framework

This section presents the design and implementation of VoxVista, a framework developed to enhance screen reading of online user comments for blind users[2]. First, we constructed a voice preference dataset capturing sighted users' perceptions and preferences regarding narration of different types of comments, resulting in a collection of voice profiles. Next, we designed LLM-based *user* and *observer* agents that leverage this voice-preference dataset to determine and assign contextually-appropriate voice profiles to different comments in real time. Finally, we evaluated VoxVista by assessing the accuracy and appropriateness of the assigned voice profiles in real-world online discussions.

4.1 Phase 1: Building a Voice Preference Dataset

We constructed a voice preference dataset capturing sighted users' perceptions regarding narration of comments. Because sighted individuals constitute the majority of contributors on public discussion platforms, modeling their expressive intentions ensures that screen readers narrate comments in a manner consistent with how they were originally intended to be perceived.

Participant Recruitment. We recruited 300 sighted participants, with a gender distribution of 167 females and 133 males. Recruitment was done through email lists and word-of-mouth referrals. Email lists were used to reach a diverse pool of active online users from different academic and professional backgrounds. The average age was 25.98 years (Median = 26, Max = 30, Min = 22). All participants met the inclusion criteria of being actively engaged in comment sections on platforms such as Reddit, Twitter, Amazon Reviews, and Canvas discussions.

Data Collection Each participant completed a 10 minute remote Google Forms survey, during which they annotated 25 comments, resulting in 7,500 annotated comment-voice pairs. The comments were sampled from real world online platforms and they belonged to one of the five categories: informational, opinionated, angry or frustrated, casual, and humorous. For each comment, the participants selected:

- A voice preference level: *Casual*, *Neutral*, or *Formal*;
- A voice tone: *Personable*, *Confident*, *Empathetic*, *Engaging*, *Witty*, or *Direct*;
- A voice source: A*I-generated*, *Own voice*, or *No preference*;
- Privacy comfort level when using personalized TTS mimicking their voice: 5-point Likert scale.

These design dimensions were informed by prior research on expressive TTS systems [43] and voice style frameworks from commercial tools such as Grammarly's voice feature[3].

Data Analysis From the 7,500 annotations, consistent quantitative patterns emerged.Across all responses, Neutral, a natural and conversational speech style, dominated with 50% of selections, followed by Casual Voice (31%) and Formal (19%). Robotic or monotonous delivery styles were intentionally not included, aligning with the interview study (Section 3) in which blind screen reader users consistently characterized flat, monotonous narration as cognitively taxing and attention-degrading during extended listening; this feedback motivates VoxVista's focus on introducing controlled variation within natural-sounding voices rather than relying on a single fixed delivery style. In terms of tone, participants most often preferred Personable and Engaging (42%), with Confident/Witty (28%) and Direct (24%) following, while highly emotional tones such as Empathetic (6%) were rarely selected, suggesting a bias toward neutral, professional delivery. Finally, for voice source, 64% of respondents preferred AI-generated voices and 36% favored using their own mimicked voice, and this coexistence of strong privacy concern alongside interest in personal voices reflects an *own voice paradox*: users value the authenticity of self-voice yet remain apprehensive about data privacy in voice-cloning systems.

4.2 Phase 2: User Profiling using LLM

VoxVista was developed using an LLM (GPT-4o [21]) with two distinct agents, each tasked with a critical function: the *user* agent analyzes the conversational flow to assign appropriate voice profiles, while the *observer* agent identifies comments of interest (i.e., entangled comments) that provide relevant context for the current voice assignment. Although voice assignment could be modeled as a supervised classification task, we adopt this formulation to enable context-aware, online adaptation in dynamic discussion threads, incorporating prior comments and user preferences.

4.2.1 ***Step 1. Creating the user agent.*** The user agent simulates how a sighted user assigns expressive voices to online comments (see Figure 2). It essentially replicates the decision-making process behind voice selection, choosing characteristics that match

[2]https://youtube.com/shorts/mfyViYcxGa8

[3]https://support.grammarly.com/hc/en-us/articles/23153676821773-Introducing-voice-features

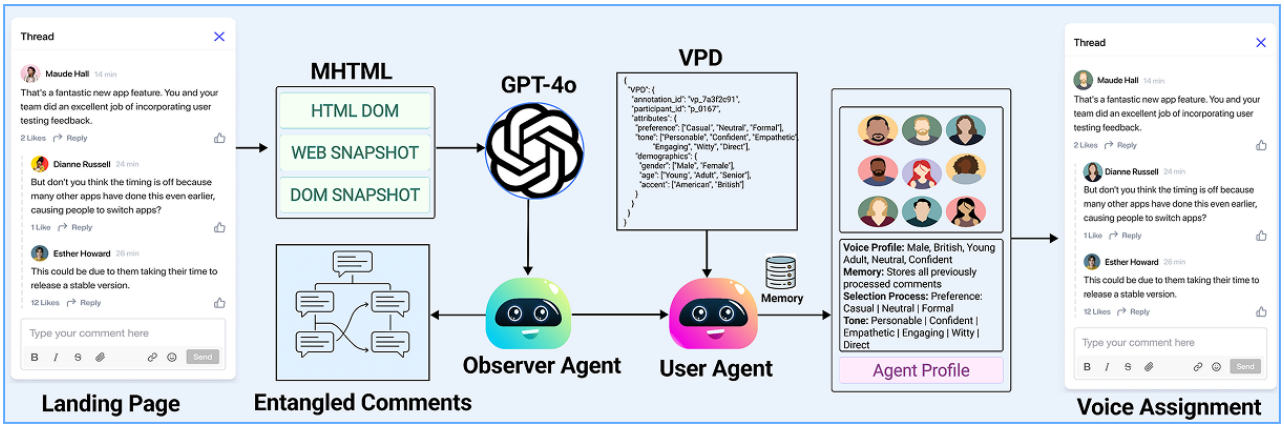

Figure 2: Overview of VoxVista Design Framework.

the intent and style of each comment. To achieve this, the user agent draws on patterns captured in the Voice Preference Dataset (Section 4.1) that contains sighted users' specified mappings between different comment types and specific voice profiles. The core constructs of the user agent are as follows:

Understanding environment state. The user agent first analyzes the content and layout of the webpage, focusing on two primary interface styles: flat view and threaded view. In the flat view, comments are presented in chronological order, facilitating linear reading of the conversation. In the threaded view, comments are nested under parent comments, helping users track discussions that span multiple subtopics. To perform this analysis, VoxVista takes a self-contained MHTML file of the webpage, which includes the raw HTML, a DOM snapshot, and a screenshot capturing layout and style. GPT-4o [21] is then used to process this input and identify the structure of the comment section, providing a detailed understanding of the environment state that guides both the user and observer agents in their tasks.

User profiling and persistent identity. The user agent assigns each author a voice profile with five attributes: *gender*, *accent*, *age*, *preference*, and *tone*. When an author is first encountered, demographic attributes (gender, accent, age) are initialized by sampling from the empirical distributions in the Voice Preference Dataset and stored in a persistent mapping keyed by the author identifier (username/user ID extracted from HTML metadata); these demographics remain fixed for that author's subsequent comments to preserve identity consistency. In contrast, expressive attributes (preference, tone) are inferred per comment from its content and conversational role, using context retrieved from entangled prior comments (Section 4.2.2), aligning selections with VPD-derived patterns for similar comments. For example, if user 'TechFan98' is assigned Male, Young Adult, American accent, all subsequent comments by 'TechFan98' retain these demographics while preference and tone adapt to each comment's role in the discussion.

4.2.2 ***Step 2. Engaging observer agent.*** After initializing the user agent, VoxVista assigns a voice profile to the first comment and stores this assignment in its long-term memory. The user agent then proceeds sequentially through the thread, assigning a voice profile to each comment. Starting from the second comment, VoxVista additionally invokes the observer agent to identify which previously processed comments are entangled with the current comment. The observer operates over the webpage structure (threaded or flat view) and semantic overlap, returning a set of contextually related prior comments that the user agent uses to inform the current voice assignment. The observer returns this set of entangled comments (and their associated user profiles) to the user agent, which uses the retrieved context to decide the current voice assignment (Figure 3).

Comment entanglement. In a threaded view, comments are often organized into sub-conversations that reflect underlying topics rather than strict reply hierarchies. On platforms such as Reddit and Stack Overflow, users frequently respond to multiple earlier comments or introduce new topics mid-thread, leading to comment entanglement. Although thread metadata captures explicit reply links, it often fails to represent these semantic overlaps. For example, in Reddit discussions on electric vehicles, a single comment may address both fast-charging infrastructure and battery degradation, creating links that cut across the rendered thread structure. Guided by ReAct prompt engineering [27], the observer component identifies comments belonging to the same sub-conversation as the comment-in-focus, by analyzing conversational content rather than relying solely on structural metadata (Figure 3).

Suppose the user agent is looking at a comment discussing the efficiency of fast-charging stations. The observer agent scans earlier comments and identifies a sub-conversation focused on charging infrastructure challenges, where multiple users debate the pros and cons of fast charging. However, due to the structure of threaded discussions, comments often respond to multiple earlier posts, leading to additional entanglements. For instance, another comment might address both fast-charging efficiency and battery degradation simultaneously. This intertwining of topics creates complex relationships between comments that span across different sub-conversations, making it challenging for the agent to isolate and evaluate comments relevant to a single sub-topic [2].

VoxVista addresses this challenge through a two-phase approach in the observer agent's reasoning. First, it identifies structural entanglements by tracing parent-child comment relationships in the

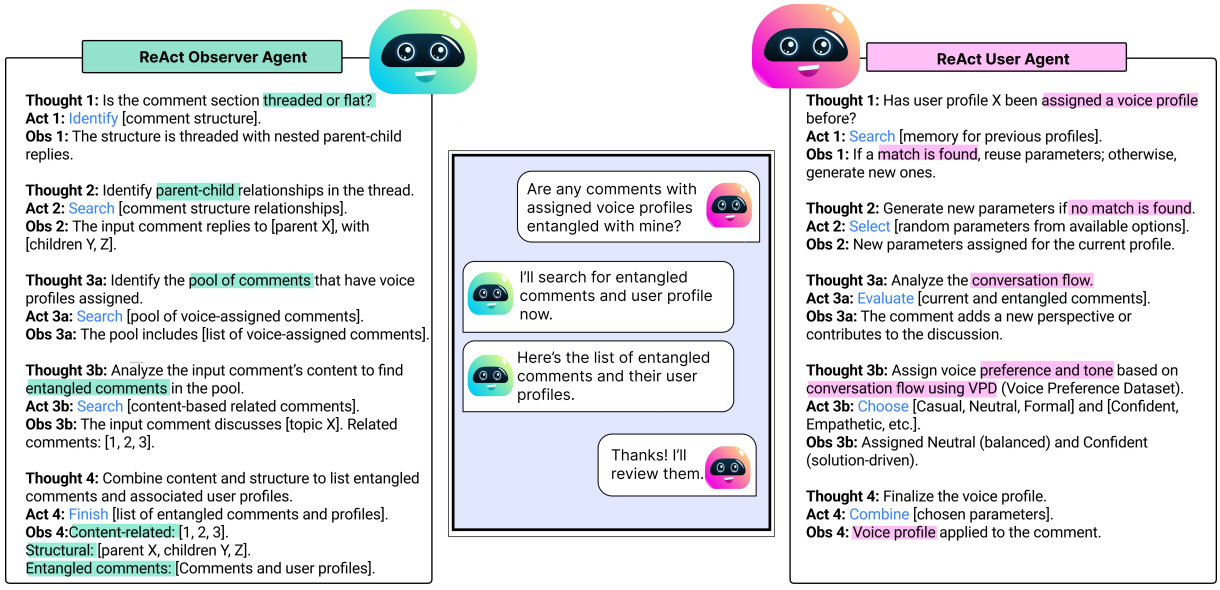

Figure 3: Observer and user agent conversation to identify comment entanglement and select appropriate voice profiles.

DOM tree. Second, it performs semantic analysis to extract all topics discussed in the current comment (e.g., fast-charging, battery degradation) and identifies comments from the pool that discuss any of these topics. When a comment spans multiple sub-conversations, the observer agent returns the union of all relevant entangled comments across these sub-topics. For example, if a comment addresses both fast-charging efficiency and battery degradation, the observer returns entangled comments from both sub-conversations. The user agent then considers this comprehensive set of entangled comments when determining voice assignment, ensuring that the assigned voice reflects the comment's multi-faceted role in the broader discussion rather than being constrained to a single sub-topic.

In a flat view scenario, the observer agent's task is simplified, as there are no hierarchical relationships. Guided by a ReAct prompt, the observer agent identifies all parent comments that are thematically entangled to the current comment by scanning previous comments for common topics.

Step 3. Voice profiling process. Once the user agent receives the list of entangled comments and associated user profiles, it leverages its long-term memory to check if the user profile linked to the current comment has previously been assigned a voice profile. If a match is found, the agent reuses the previously assigned speech parameters. If no match is found, the agent samples these parameters from the VPD empirical distribution, building a new persona for the current comment's author (Figure 3).

Next, the user agent sets two remaining profile parameters: voice preference and tone, from its action space. Towards this, the agent leverages its understanding of the environment to evaluate how the current comment contextually fits into the conversational flow, i.e., whether it is built on previous points, introduces a new perspective, or reinforces an existing consensus within the discourse. For example, in a discussion about the benefits and drawbacks of remote work, the user agent must determine the appropriate voice for a comment that addresses both perspectives:

- Comment 1: "*Remote work allows for a better work-life balance and saves commuting time.*" (**Assigned Voice: Casual and Personable**)
- Comment 2: "*While that's true, not everyone is suited for remote work; some people struggle with isolation.*" (**Assigned Voice: Neutral and Empathetic**)
- Comment 3 (new comment): "*I agree that remote work can improve productivity, but companies should also provide support for mental health to address isolation.*" (**Assigned Voice: ?**)

The user agent, guided by a ReAct prompt, assesses the above conversation flow as follows. It recognizes that Comment 1 emphasizes the positive aspects of remote work, while Comment 2 highlights concerns about isolation. Understanding that Comment 3 addresses both points, i.e., acknowledging the benefits while also recognizing the challenges, the user agent taps into its understanding of how sighted users assign voice preference and tone to comments, and then assigns a *Neutral* voice preference to project a balanced and thoughtful tone. It next selects a *Confident* voice tone to reflect the solution-oriented nature of the comment.

Injecting "profile voices" to user comments. Once the user agent determines voice profiles to all comments, VoxVista consults a custom dataset containing 216 voice options (comprising 3 voice preference choices x 6 voice tone choices x 3 age ranges x 2 accents x 2 genders) built using Google Custom Voice TTS[4]. Specifically, VoxVista queries this dataset to identify a matching voice based on the combination of voice profile parameters and embeds it directly into the HTML structure of the comment (see Figure 2).

[4]https://cloud.google.com/text-to-speech/custom-voice/docs

4.3 Assessment

We assessed both agents' performance on 100 websites (25 each from Reddit, Amazon reviews, X, and Canvas). Seven annotators with expertise in online discourse analysis were recruited and trained over two sessions on comment entanglement (structural and semantic relationships) and voice profiling (preference and tone dimensions). They completed practice annotations on 10 sample webpages to establish inter-rater reliability. For the observer agent, annotators identified all entangled comments from a seed comment by considering thread structure and semantic overlap. For the user agent, annotators analyzed conversational excerpts of 3-5 contextually related comments to understand flow before annotating appropriate preference and tone. Excerpts were necessary because voice assignment depends on conversational context, where a comment's relationship to prior discussion determines whether it should be voiced as casual versus formal or confident versus empathetic.

We could not use a held-out test set from the Voice Preference Dataset (Section 4.1) for two reasons. First, the VPD contains isolated comments without conversational context or entanglement information, whereas our evaluation requires assessing the agents' ability to process multi-comment discussions. Second, the VPD was collected to train the agents through few-shot learning; using the same distribution for testing would not assess generalization to novel conversational patterns encountered in real-world forums.

Observer Agent Evaluation. For each seed comment, annotators identified all entangled comments within the discussion thread, creating a ground truth set. The observer agent's predictions were compared against this ground truth using standard information retrieval metrics. Precision measures the proportion of agent-identified comments that were truly entangled, while Recall measures the proportion of actual entangled comments that the agent successfully identified. The observer agent achieved a mean Precision of 0.86, mean Recall of 0.82, and mean F1-score of 0.839 across 100 webpages, indicating strong performance in identifying relevant comments while maintaining high coverage.

User Agent Evaluation. For voice profile assignment, we evaluated the two parameters separately using multi-class classification metrics. Voice preference has 3 classes (Casual, Neutral, Formal) and voice tone has 6 classes (Personable, Confident, Empathetic, Engaging, Witty, Direct). We computed precision and recall for each parameter, then calculated macro-averaged F1-scores to account for class imbalance. The user agent achieved Precision of 0.85, Recall of 0.81, and F1-score of 0.83 for voice preference assignment. For voice tone assignment, the agent achieved Precision of 0.81, Recall of 0.77, and F1-score of 0.79.

Inter-annotator agreement, measured using Fleiss' kappa, was 0.74 for entanglement identification (substantial agreement) and 0.68 for voice assignment (substantial agreement), confirming the reliability of our ground-truth annotations. To relate inter-annotator agreement to agent performance, we evaluated annotator-to-annotator agreement by treating one annotator's labels as ground truth, averaged across annotator pairs and webpages. For entanglement identification, annotators achieved a mean F1-score of 0.86 (Precision = 0.88, Recall = 0.84). For voice profile assignment, agreement was lower, with a mean F1-score of 0.80 (Precision = 0.83, Recall = 0.78). These results indicate that both the observer agent (F1 = 0.839) and user agent (F1 = 0.81) perform within the range of expert human agreement on these tasks.

5 VoxVista Evaluation

We conducted an IRB-approved study to evaluate the effectiveness of VoxVista. We recruited 20 BVI participants[5] (8 female, 12 male), distinct from the interview study, with an average age of 32.1 years (Median = 31, SD = 6.83, Range = 23-43), through email lists and snowball sampling. Inclusion criteria required participants to be proficient with screen readers and familiar with online comments.

Design. In a within-subject setup, the participants were asked to interact with each of the four platforms (Reddit, Amazon reviews, X, and Canvas) under three distinct study conditions: (i) Screen Reader Robotic voice (i.e. Windows OneCore voice); (ii) Screen Reader Natural voice (Microsoft Speech API Version 5), and (iii) VoxVista generated voice. We selected the same websites used earlier for evaluating VoxVista agents, and chose NVDA as the screen reader based on participants' preferences gathered during recruitment.

The study task required participants to navigate user comments for up to 10 minutes while untangling the conversational flow. This task, informed by prior research [32], was designed to replicate realistic scenarios capturing how most users browse comment sections. The study included 12 combinations (3 conditions × 4 platforms), with 8 webpages without voice profiles and 4 webpages with voice profiles and total task time was 120 minutes. The assignment of platform webpages to study conditions, and the order of conditions, was counterbalanced using the Latin Square method [6].

Apparatus. The study was conducted using a Lenovo ThinkPad laptop. For the first two study conditions, the NVDA screen reader was installed under the Microsoft Windows operating system. For the third study condition, VoxVista's user interface was created by processing the webpage through VoxVista, embedding voice profiles into each user comment. A custom screen reader interface was developed, offering basic navigation functionality through TAB and arrow keys to focus on and move between elements. The selection of keyboard shortcuts was guided by participant feedback gathered during the recruitment process, ensuring negligible learning curve. An external QWERTY desktop keyboard was also connected to the laptop, as all participants indicated familiarity with the standard keyboard during the recruitment process.

Procedure. The experimenter initiated the study by obtaining informed consent from the participants and providing an overview of the study's objectives. Participants were then given practice time with VoxVista to ensure familiarity with the system. The experimenter next instructed participants to complete tasks in a counterbalanced order. For the VoxVista condition, NVDA was turned off, while for the other conditions, NVDA was activated. The study commenced when the participant reached the first user comment, with attention focused solely on the user comments, disregarding the influence of other HTML elements on the webpage. After completing the task across all four platforms in each condition, participants were administered a VECS (Voice Experience Comparative Study) likert survey (see Figure 4). At the end of the study, the

[5]Participant demographics are available at: https://github.com/accessodu/VoxVista.git

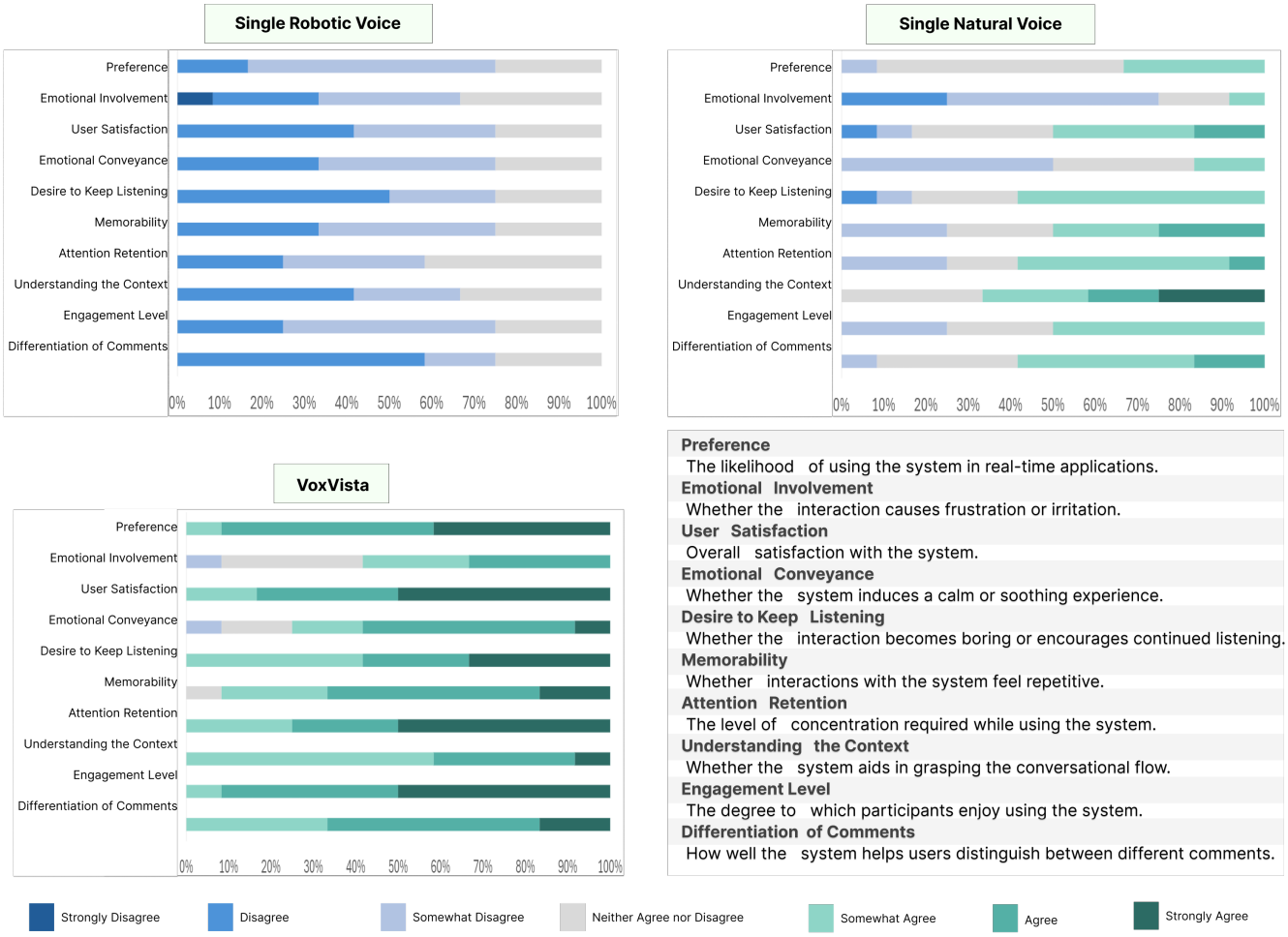

Figure 4: VECS ratings comparison for single robotic voice, single natural voice, and VoxVista.

experimenter collected subjective feedback from the participants via exit interviews.

Results. The VECS questionnaire comprises ten 7-point Likert scale questions (1-strongly disagree, 7-strongly agree). This survey is an adaptation of the Subjective Assessment of Speech System Interfaces (SASSI) questionnaire [17], designed specifically to test our hypothesis that using multiple tailored voice profiles for user comments enhances content engagement. The focus was on factors like likeability, annoyance, and cognitive demand to measure the system's efficacy in delivering an engaging auditory experience.

Figure 4 shows a clear shift in VECS ratings toward the positive end of the scale for VoxVista compared to both single-voice baselines. In the single robotic voice condition, responses for most dimensions are concentrated in the mid-to-lower ranges, indicating reduced preference, weaker engagement, and higher perceived effort. The single natural voice baseline improves substantially across nearly all items, with many participants selecting higher ratings for satisfaction, desire to keep listening, and emotional conveyance. However, VoxVista produces the most consistent improvement overall: across *Preference*, *User Satisfaction*, *Engagement Level*, *Desire to Keep Listening*, *Attention Retention*, and especially *Differentiation of Comments*, where responses are heavily concentrated in the highest rating bins, suggesting that multi-voice rendering makes comment threads easier to follow. Importantly, the item-level breakdown revealed a nuanced trade-off: *Understanding the Context* is slightly higher for the single natural voice condition than for VoxVista, suggesting that the familiarity and continuity of a single narrator may sometimes support perceived conversational coherence, whereas frequent voice switching can introduce a short adaptation cost even when it improves speaker separation and engagement.

For statistical analysis, we computed an aggregate VECS score for each participant in each condition by averaging their responses across all ten items. This approach is justified because the items measure related dimensions of the same construct (auditory user experience). The aggregate VECS scores for the *VoxVista* condition were significantly higher than both *natural voice screen reader* and *robotic voice screen reader* conditions. We conducted a one-way repeated measures ANOVA with condition (three levels: robotic voice, natural voice, VoxVista) as the within-subjects factor and aggregate VECS score as the dependent variable. The effect of condition was

statistically significant ($F(2, 22) = 39.28$, $p < 0.001$), indicating that voice configuration significantly impacted user experience. A post-hoc Tukey's HSD test revealed that the differences in aggregate VECS scores were statistically significant across all pairwise comparisons: (i) *single robotic voice* vs. *VoxVista* ($Q = 19.97$, $p < 0.001$); (ii) *single natural voice* vs. *VoxVista* ($Q = 10.62$, $p < 0.001$); and (iii) *single robotic voice* vs. *single natural voice* ($Q = 9.35$, $p < 0.001$).

Exit Interviews. A majority of participants (90%) expressed a strong desire to continue using VoxVista, describing the experience of listening to user comments as akin to 'an audiobook with a narrative style'. A substantial portion (75%) reported enhanced engagement with VoxVista, expressing excitement as they followed the unfolding conversation, and preferring to hear multiple distinct voices rather than a single one. However, 25% of participants were undecided, noting that they also appreciated the consistency of a single voice. All participants (100%) indicated that with VoxVista, they were able to differentiate between comments and comprehend the context of the conversation easily. Moreover, they acknowledged that understanding the flow required more effort and attention when using single natural and robotic voices. Regarding emotional reliability, 35% of participants noted that they could experience emotions through natural voice and sometimes through robotic voice as well, but emphasized that this depended on the context of the comment. For platforms like Amazon, where the content is predominantly informational, a real emotional connection was less relevant. They indicated that for more genuine emotional engagement, the conversation needed to involve dialog they could personally relate to, such as those attached to news articles.

6 Discussion

The user study revealed that VoxVista significantly enhances engagement and experience when interacting with user comments, extending prior research on single-voice TTS [10], by demonstrating that voice diversity is critical for multi-party conversational contexts. Our findings align with broader HCI research showing that voice characteristics influence interaction dynamics [10, 23, 33], while providing empirical evidence that the transition from monotone to contextually-aware multi-voice narration meaningfully improves comprehension, engagement, and willingness to continue listening. However, our study also had the following limitations.

Limitations. A limitation was our focus on only English-language platforms. In real-world scenarios, users engage in multilingual discourse depending on the platform. Expanding our approach to support multilingual voice profiles is a crucial avenue for future research. Additionally, our VPD dataset, which trains agents for voice profiling, is currently constrained to four platforms. However, voice profile choices may be influenced by personal censorship, depending on the audience accessing the comments [43]. Therefore, we aim to expand our dataset, as it is still in its early stages.

Another limitation was that the participants were not allowed to adjust the listening rate of the screen reader, which was fixed at 140 WPM. This restriction was applied to avoid confounding variables that would require separate evaluation [7, 8], which we plan to explore in future work. In this paper, we regarded this to be a 'trade-off' between maintaining a consistent listening rate and transitioning from a single-voice to a multi-voice system. Also, while our within-subject design mitigates between-participant differences, factors such as age and prior screen reader or technology experience may still influence adaptation to voice switching and perceived cognitive demand; our sample was not powered to test these moderators, and we will address this in future studies.

Our evaluation of LLMs focused primarily on comparing the outputs with baseline human annotators, without assessing more complex aspects such as the agent's self-awareness or its ability to exhibit sophisticated behaviors [42]. In future work, we aim to develop a more comprehensive evaluation framework, addressing the limitations of existing approaches. Additionally, our study on threaded views was limited to platforms with a simple parent-child comment structure, and we did not explore more complex, multi-layered conversation threads. Expanding the scope to include such complexities is another key future research direction.

Finally, VoxVista's voice inventory varies not only in the modeled expressive dimensions (voice preference and tone), but also in demographic dimensions such as accent and gender. Because these attributes can increase perceptual distinctiveness and help listeners separate speakers, our VECS gains may partially reflect demographic variation rather than the preference/tone assignments. Thus, while our study evaluates VoxVista as an end-to-end multi-voice system, it does not isolate the causal contribution of each profile attribute. To disentangle these effects, future work will conduct factorial ablations comparing: (i) a single natural voice, (ii) multi-voice with only accent/gender variation, (iii) multi-voice with only preference/tone variation, and (iv) the full VoxVista configuration. This setup will estimate the added benefit of preference/tone modeling beyond speaker differentiation from demographic variety.

Long Term Memory and Personalization. In VoxVista, the long-term memory of the agents was limited to age, gender, and accent, which were randomly assigned by the user agent. In future work, we plan to expand to include voice preference and tone, enabling the agent to create a more comprehensive user persona. Instead of having a single agent assign profiles to all comments, we envision deploying multiple agents across different platforms and webpages. These agents will learn to assign a diverse set of preferences and tones, gradually developing their ability to build complete and personalized voice profiles independently.

7 Conclusion

Blind users relying on screen readers for their daily interaction with user comments are often limited to a single robotic or monotonic voice, lacking expression and diversity. To address this issue and test the hypothesis that transitioning from a single-voice screen reader to an alternative multi-voice experience enhances user engagement with online user-generated content such as comments and discussions, we introduced VoxVista, an automatic voice-profile assignment system powered by an LLM agent. This agent understands the context of comments as part of a larger discourse and dynamically assigns profile voices to comments. In a user study with 20 participants, we found that VoxVista significantly improved the user experience, making listening to comments more engaging and dynamic. Future work will focus on incorporating multilingual LLM capabilities, developing in-depth evaluation frameworks, and extending VoxVista to include long-term memory personalization.

References

[1] Deepak Agarwal, Bee-Chung Chen, and Bo Pang. 2011. Personalized recommendation of user comments via factor models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*. 571–582.

[2] Anand Ravi Aiyer, IV Ramakrishnan, and Vikas Ashok. 2023. Taming Entangled Accessibility Forum Threads for Efficient Screen Reading. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 65–76.

[3] Ben Alderson-Day and Charles Fernyhough. 2015. Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin* 141, 5 (2015), 931–965.

[4] Chieko Asakawa, Hironobu Takagi, Shuichi Ino, and Tohru Ifukube. 2003. Maximum listening speeds for the blind. In *Proceedings of the 2003 International Conference on Auditory Display*. 276–279.

[5] Marialena Barouti, Konstantinos Papadopoulos, and Georgios Kouroupetroglou. 2013. Synthetic and natural speech intelligibility in individuals with visual impairments: Effects of experience and presentation rate. In *Assistive Technology: From Research to Practice*. IOS Press, 695–701.

[6] James V. Bradley. 1958. Complete Counterbalancing of Immediate Sequential Effects in a Latin Square Design. *J. Amer. Statist. Assoc.* 53, 282 (1958), 525–528. arXiv:https://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1958.10501456 doi:10.1080/01621459.1958.10501456

[7] 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*. 1–12.

[8] Danielle Bragg, Katharina Reinecke, and Richard E Ladner. 2021. Expanding a large inclusive study of human listening rates. *ACM Transactions on Accessible Computing (TACCESS)* 14, 3 (2021), 1–26.

[9] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[10] Julia Cambre and Chinmay Kulkarni. 2019. One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–19.

[11] Philip R Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R Cowan. 2019. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st international conference on human-computer interaction with mobile devices and services*. 1–12.

[12] Md Javedul Ferdous, Akshay Kolgar Nayak, Yash Prakash, Nithiya Venkatraman, Sampath Jayarathna, Hae-Na Lee, and Vikas Ashok. 2025. Understanding Online Discussion Experiences of Blind Screen Reader Users. *International Journal of Human–Computer Interaction* (2025), 1–31.

[13] Mukta Gahlawat, Amita Malik, and Poonam Bansal. 2014. Natural speech synthesizer for blind persons using hybrid approach. *Procedia computer science* 41 (2014), 83–88.

[14] João Guerreiro and Daniel Gonçalves. 2014. Text-to-speeches: evaluating the perception of concurrent speech by blind people. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. 169–176.

[15] 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*. 3–11.

[16] Dominique Heinbach, Marc Ziegele, and Oliver Quiring. 2018. Sleeper effect from below: Long-term effects of source credibility and user comments on the persuasiveness of news articles. *New media & society* 20, 12 (2018), 4765–4786.

[17] Kate S Hone and Robert Graham. 2001. Subjective assessment of speech-system interface usability. In *Seventh European Conference on Speech Communication and Technology*.

[18] Angela King and Barbara Dodd. 2020. The effects of auditory cues on blind users' comprehension of digital content. *Journal of Visual Impairment & Blindness* 114, 3 (2020), 223–234.

[19] Eun-Ju Lee and Yoon Jae Jang. 2010. What do others' reactions to news on internet portal sites tell us? Effects of presentation format and readers' need for cognition on reality perception. *Communication research* 37, 6 (2010), 825–846.

[20] Fabrice Maurel, Gaël Dias, Stéphane Ferrari, Judith-Jeyafreeda Andrew, and Emmanuel Giguet. 2019. Concurrent speech synthesis to improve document first glance for the blind. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 3. IEEE, 10–17.

[21] Agnieszka Mensfelt, Kostas Stathis, and Vince Trencsenyi. 2024. Logic-Enhanced Language Model Agents for Trustworthy Social Simulations. *arXiv preprint arXiv:2408.16081* (2024).

[22] Teresa K Naab and Constanze Küchler. 2022. Content analysis in the research field of online user comments. In *Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft–Standardized Content Analysis in Communication Research: Ein Handbuch-A Handbook*. Springer Fachmedien Wiesbaden Wiesbaden, 441–450.

[23] CI Nass. 2005. Wired for speech: How voice activates and advances the human-computer relationship.

[24] Jonathan E. Peelle. 2018. Speech comprehension: Stimulating, engaging, and challenging the brain. *Current Opinion in Behavioral Sciences* 24 (2018), 30–35.

[25] M. Kathleen Pichora-Fuller and et al. 2016. Hearing, cognition, and healthy aging: Social and public health implications. *Ear and Hearing* 37, Suppl 1 (2016), S1–S7.

[26] Monika Podsiadło and Shweta Chahar. 2016. Text-to-speech for individuals with vision loss: a user study. *Interspeech 2016* (2016), 347–351.

[27] Fina Polat, Ilaria Tiddi, and Paul Groth. 2024. Testing Prompt Engineering Methods for Knowledge Extraction from Text. *Semantic Web. Under Review* (2024).

[28] Johnny Saldaña. 2015. *The coding manual for qualitative researchers*. Sage.

[29] Patric R Spence, Kenneth Lachlan, Timothy Sellnow, Robert G Rice, and Henry Seeger. 2017. That is so gross and I have to post about it: Exemplification effects and user comments on a news story. *Southern Communication Journal* 82, 1 (2017), 27–37.

[30] Natalie Jomini Stroud, Emily Van Duyn, and Cynthia Peacock. 2016. News commenters and news comment readers. *Engaging News Project* 21 (2016), 1–21.

[31] Mohan Sunkara, Akshay Kolgar Nayak, Sandeep Kalari, Yash Prakash, Sampath Jayarathna, Hae-Na Lee, and Vikas Ashok. 2025. QuickQue: Enabling Quick Access to Information in User Reviews for Screen Reader Users. In *Proceedings of the 22nd International Web for All Conference*. 22–24.

[32] Mohan Sunkara, Yash Prakash, Hae-Na Lee, Sampath Jayarathna, and Vikas Ashok. 2023. Enabling Customization of Discussion Forums for Blind Users. *Proceedings of the ACM on Human-Computer Interaction* 7, EICS (2023), 1–20.

[33] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.

[34] Nithiya Venkatraman, Anand Aiyer, Yash Prakash, and Vikas Ashok. 2024. You Shall Know a Forum by the Words they Keep: Analyzing Language Use in Accessibility Forums for Blind Users. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*. 230–238.

[35] Nithiya Venkatraman, Anand Ravi Aiyer, Yash Prakash, Sampath Jayarathna, Hae-Na Lee, and Vikas Ashok. 2025. Characterizing Language Use in Online Accessibility Discussion Forums. *ACM Transactions on the Web* (2025).

[36] T Franklin Waddell. 2020. The authentic (and angry) audience: How comment authenticity and sentiment impact news evaluation. *Digital Journalism* 8, 2 (2020), 249–266.

[37] Joseph B Walther, David DeAndrea, Jinsuk Kim, and James C Anthony. 2010. The influence of online comments on perceptions of antimarijuana public service announcements on YouTube. *Human communication research* 36, 4 (2010), 469–492.

[38] Yixue Wang. 2021. Comment section personalization: Algorithmic, interface, and interaction design. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. 84–88.

[39] Gregory Weinstein. 2019. Hearing Through Their Ears: Developing Inclusive Research Methods to Co-Create with Blind Participants. In *Ethnographic Praxis in Industry Conference Proceedings*, Vol. 2019. Wiley Online Library, 88–104.

[40] Spencer Williams and Gary Hsieh. 2021. The effects of user comments on science news engagement. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–29.

[41] Stephan Winter and Nicole C Krämer. 2016. Who's right: The author or the audience? Effects of user comments and ratings on the perception of online science articles. *Communications* 41, 3 (2016), 339–360.

[42] Qiuejie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Yuejie Zhang, Rui Feng, and Shang Gao. 2024. Human Simulacra: A Step toward the Personification of Large Language Models. *arXiv preprint arXiv:2402.18180* (2024).

[43] Lotus Zhang, Lucy Jiang, Nicole Washington, Augustina Ao Liu, Jingyao Shao, Adam Fourney, Meredith Ringel Morris, and Leah Findlater. 2021. Social media through voice: Synthesized voice qualities and self-presentation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[44] Marc Ziegele, Christina Koehler, and Mathias Weber. 2018. Socially destructive? Effects of negative and hateful user comments on readers' donation behavior toward refugees and homeless persons. *Journal of Broadcasting & Electronic Media* 62, 4 (2018), 636–653.