Toward Effective Communication of AI-Based Decisions in Assistive Tools: Conveying Confidence or Doubt to People with Visual Impairments at Accelerated Speech

Taslima Akter
taslima@uci.edu
University of California Irvine
Irvine, California, USA

Manohar Swaminathan
manohar.swaminathan@microsoft.com
Microsoft Research India
Bengalore, India

Apu Kapadia
kapadia@indiana.edu
Indiana University Bloomington
Bloomington Indiana, USA

ABSTRACT

AI-based assistive technologies can mischaracterize information presented to people with visual impairments (PVIs), e.g., by misinterpreting someone's facial expressions. To improve their trustworthiness, algorithmic decisions can be communicated more effectively by conveying their level of confidence to PVIs. Since assistive feedback is typically provided through accelerated audio for PVIs, this work explores how the tone of the audio output can be manipulated to convey confidence or doubt at higher speeds for short sentences. We conducted two online surveys with PVI (n = 151) and sighted (n = 170) participants. We found that PVI participants perceived confidence and doubt for short sentences more accurately at up to 1.5x speedups, whereas, sighted participants perceived them at up to 2x speedups. Other factors such as preferred speedups are also associated with better perception. These results demonstrate the potential for conveying the confidence level of AI-based decisions by choosing the appropriate speedup rates.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in accessibility**.

KEYWORDS

confidence, doubt, accelerated speech, visually impaired, assistive tools

ACM Reference Format:
Taslima Akter, Manohar Swaminathan, and Apu Kapadia. 2024. Toward Effective Communication of AI-Based Decisions in Assistive Tools: Conveying Confidence or Doubt to People with Visual Impairments at Accelerated Speech. In *W4A '24: Web for All Conference, May 13–14, 2024, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

People with visual impairments (PVIs) have a long history of using audio interaction-based assistive technologies such as screen readers [5, 7, 10] and optical character recognition (OCR) devices [6, 9]. PVIs are also adopting voice-based personal assistants or conversational agents (e.g., Amazon Echo, Google Home) that support non-visual interaction [4, 8]. Moreover, audio description (AD) enhances the experience of PVIs when accessing audiovisual content (e.g., visual images of theater, television, movies, and other art forms) [41] and audiobooks are creating an environment for PVIs to learn and develop independent learning skills [78]. Camera-based assistive technologies, too, are now presenting the visual world as a form of audible experience to people with visual impairments [3, 11, 12]. Such systems analyze the visual data around PVIs and provide audio feedback to help them with scene description [12], and recognizing faces [11] and people's behaviors and emotions [12]. Furthermore, with the advancement of large language models, products such as ChatGpt [15] are now able to provide real-time explanations of images to users. This capability can be utilized to generate descriptions of visual scenes and convey them through audio.

Despite the promise of providing audio-based feedback in the context of scene description, prior studies have reported the concerns of PVIs, as well as people around them (bystanders), about misrepresentations caused by camera-based assistive technologies [20, 21]. Research participants with visual impairments reported hesitance to receive socially-biased and subjective information about bystanders to avoid embarrassing scenarios if the device's AI-based judgments were inaccurate. Bystander participants, too, were concerned about being misrepresented by automated technologies [21]. Bennett et al. discussed how Black, Indigenous, Person of Color (BIPOC), non-binary, and/or transgender screen reader users shared concerns related to the accuracy and ethical deployment of AI-based image descriptors [26]. Some participants were frustrated by the binary gender classifications of these AI-based systems and described how they were negatively impacted, particularly those who identified as non-binary or transgender. As a result, PVIs still lack trust in AI-based assistive technologies and often prefer human assistance, as opposed to AI-based systems, because of their inaccurate and inconsistent responses [19, 69].

Research in the area of "explainable AI" has sought to improve people's trust in AI-based decisions by attempting to explain algorithmic decision-making [36, 43]. It is possible to heighten user awareness of AI limitations by indicating when the certainty of suggestions made by the algorithm is low [66]. Prior works explore how the framing (positive vs. negative) of AI-based decisions affects people's trust in the system [56, 69]. Macleod et al. recommended phrasing the automatically-generated captions in a way that indicates the possibility of the caption being wrong (negative

W4A '24, May 13–14, 2024, Singapore

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

framing) [69]. However, it is also possible to explain AI-based decisions through audio because the tone and inflection of the voice can influence how the information is perceived. We believe that investigating how to communicate – through speech – the limitations of algorithmic decisions to PVIs would help them make better decisions from AI-based suggestions. Prior work describes the importance of voice-based cues when PVIs assess trustworthiness [76, 77]. Therefore, conveying the level of confidence in AI-based determinations as audio feedback will ultimately improve the trustworthiness and social acceptability of AI-based assistive technologies among PVIs [20].

Humans understand various emotive features from speech including confidence or doubt. Prior work has found that listeners can infer the confidence of a speaker through the tone of their answers, their use of fillers, and their response latency [30, 35]. For example, rising intonation and longer latency can be an indication of a low confidence level of the speaker [30]. However, people with visual impairments have higher listening speedup rates compared to their sighted peers [27, 28, 73, 102]. Prior works also reported that accelerating video playback enables people to browse videos efficiently, thus saving time, having more control over the contents, and consuming more content [63, 65]. However, it is possible that when consuming content or interacting with screen readers in accelerated speech the perception of prosodical and emotional cues from speech might be different than at regular speed. For example, Choi et al. found that PVIs considered the voice of conversational agents more machine-like at higher speeds compared to the default speed [32]. Therefore, as a first step in effectively communicating the confidence level of AI-based decisions, it is important to understand how people with visual impairments perceive confidence and doubt from the voice of the speaker in accelerated speech.

Therefore, in this paper, we focused on the following research question: *How do people (whether visually impaired or sighted) perceive the confidence level of the speaker in accelerated speech?* To answer this question, we conducted two online surveys with 151 visually-impaired participants and 170 sighted participants, examining their perceptions of confidence and doubt from speech played at four different playback speeds. Participants completed a between-subjects survey instrument with their responses to the audio recordings of three sentences played at different speeds (default, 1.5x, 2x, and 2.5x). Our stimuli contained audio recordings of six native English actors who were instructed to convey either "confidence" or "doubt". We conducted both quantitative and qualitative analyses to understand how the perception of confidence and doubt vary for visually impaired and sighted participants, and what factors influence their perceptions.

Our findings suggest that, overall, sighted people are able to understand the confidence and doubt conveyed by the speakers at up to a 2x playback speed, although their performance drops significantly at 2.5x playback speed. We also observed that PVIs perceived confidence and doubt conveyed by speakers more correctly up to a 1.5x playback speed. We also found that listeners characteristics influence the perception of confidence and doubt. For example, visually impaired participants who usually preferred to speed up their audio or video content were able to perceive doubt comparatively better than visually impaired participants who usually do not speed up their media. We also observed that visually impaired participants who were advanced screen reader users perceived confidence more accurately compared to the participants with intermediate screen reader proficiency. Our findings have important implications for understanding the appropriate speed levels for conveying confidence (or doubt) in the algorithms of assistive systems and how to preserve the speech characteristics or non-verbal cues to adequately convey confidence (or doubt) in accelerated speech.

2 RELATED WORK

2.1 Speech Interaction-based Assistive Technology

Speech-based interaction is commonly used to support people with visual impairments (PVIs). Screen readers (e.g., JAWS [7], NVDA [10], TalkBack [13], and VoiceOver [5]) give audio output to navigate interfaces and access text by converting interfaces and digital text into spoken text. By providing hands-free interaction, speech input-based applications are also gaining popularity. With these, users can engage in a dialog with an intelligent agent that translates user commands to perform appropriate actions [23, 24]. Recently, voice-based smart personal assistants (e.g., Amazon Alexa, Google Home, Siri) have also become popular among PVIs. Prior research investigated the benefits and concerns PVIs have while using such systems in their daily lives [17, 29, 32, 90]. Given the extensive use of speech-based assistive technologies by PVIs, researchers have focused on how the listening abilities of PVIs can affect their experiences with such systems [17, 29, 32]. For example, Abdolrahmani et al. reported that PVIs were frustrated because the speed of information presented by voice assistants could not be controlled easily and they had to interact at a slower speed than expected [17].

2.2 Human Comprehension of Accelerated Speech

The average human typically speaks at about 150 words per minute or 9 syllables per second [2, 86]. In terms of listening and comprehension, prior works suggest that speech can be accelerated to nearly 250 words per minute before audio comprehension starts to worsen [83]. Many studies have found that PVIs can understand speech at faster rates [22, 73, 92, 102, 103] than sighted people. PVIs can comprehend at least 50 percent of the information at 500 words per minute [22, 32], which is around three times faster than the average speed of spoken English [2]. McCarthy et al. found that blind people generally speed up the audio output as they become experts with screen readers [71]. In earlier research, PVIs often performed better than sighted people on auditory tasks [28, 73]. In the context of musical abilities, Gougoux et al. reported that people who became blind at an early age are better at identifying relative pitch than sighted peers [46]. Blind people are more likely to have perfect pitch and the ability to identify absolute sound frequencies [50, 94] and have superior ability at localizing sound [93, 104].

Prior works also have reported that the use of higher playback speed is increasing [1, 14, 38] and it accelerates the consumption of digital information [48, 63, 74]. Duan and Chen observed that the majority of college students prefer higher playback speed while watching streaming dramas [38]. Nagahama and Morita found that viewing videos at accelerated speeds is more effective for learning than viewing at normal speed [74]. Prior works also explored the intelligibility of speech produced by different mechanisms, including natural speech with and without pauses and synthesized speech [25, 73, 80]. Papadopoulos et al. observed higher comprehensibility of synthesized speech among blind users, most probably because of their experience with screen readers [80, 81]. They found that both blind and sighted users have similar understanding of natural speech and synthesized speech [81].

The outstanding listening ability of PVIs has encouraged research on how to support their digital information acquisition. Abdolrahmani et al. recommended including customized voice output settings (e.g., speech rate, clarity, and intensity) in voice-based personal assistants, depending on context [17]. Choi et al. explored how faster speech rates affected the experience of conversational agents by PVIs [32]. They found that although faster speech rates often make conversation machine-like, PVIs sought more human-like communication with conversational agents. Therefore, more research is needed to understand how PVIs perceive embedded information (e.g., confidence or emotion) in speech at different speeds.

2.3 Perception of Confidence in Speech

Humans have a unique quality to perceive the expression of confidence and uncertainty in speech through verbal cues (e.g., linguistic phrases and constructions) [30, 37, 55, 89, 100] or non-verbal cues (e.g., tone of voice, facial expression) [35]. Brennan and Williams found that listeners can reasonably guess speakers' confidence and called it "feeling of another's knowing" (FOAK) [30]. According to them, FOAK was affected by the intonation of answers, usage of fillers, and the latency to response (e.g., a rising intonation and longer latencies often accompanied by lower FOAK ratings by listeners) [30]. Swerts and Krahmer found that listeners can make somewhat better assumptions of a speaker's level of confidence when they have access to both visual and auditory cues than only one of them [59]. Rowland referred to words children sometimes use (e.g., "about" or "maybe") as "hedges," to convey uncertainty to listeners [95]. Prior works found that listeners can recognize the confidence of speakers better with the integration of congruent linguistic and vocal cues in certain situations [54, 84, 85]. For example, Jiang and Pell revealed that an initial linguistic phrase with a congruent confident voice (e.g., "I am positive he can access the building") has a better recognition rate than the ratings for vocal cues only [54]. Moreover, linguistic information could be useful when prosodic decoding mechanisms were impaired (e.g., patients with Parkinson's disease) [72]. However, other studies have discovered that in certain experimental conditions non-verbal cues are more effective than linguistic phrases to convey the speaker's intended confidence levels [97, 105]. Overuse of linguistic phrases communicating a high level of confidence (e.g., "I am very/absolutely/so certain") can negatively impact the perceived confidence rating and listeners tend to rely more on nonverbal cues in such situations [57]. Although several studies have explored how people perceive the confidence level of the speaker based on different prosodic and lexical cues, more work is needed to understand how speech at accelerated speeds can affect the perception of confidence. Our work sheds light on such perceptual differences based on varying speeds by studying sighted and visually impaired listeners.

3 METHODS

3.1 Stimuli and Speedup Selection

In this study, we used the confidence dataset collected by Jiang and Pell [55]. The dataset contains 23 short semantically neutral sentences (no linguistic markers that might indicate the speaker's level of confidence) that could appear in daily interactions. Each sentence was produced by six native English speakers with three confidence levels: 1) confident, 2) doubt, and 3) neutral. From the 23 sentences, we randomly selected three nine-syllable sentences. The sentences we considered were: "He is on a new medication," "He is the right person to do this," and "He will visit his parents after." Our initial stimuli set, therefore, contained 54 audios (three sentences recorded by six actors with three confidence levels).

In this study, we considered four different playback speedups (default, 1.5x, 2x, and 2.5x speeds).[1] Therefore, the pilot stimuli set contained 216 audios (54 audios from each speedup). While increasing the playback speed, the pitch was preserved across different speedups for each audio file. We also normalized the volume level of the audio files.

3.2 Measurements (Dependent variables)

To measure the confidence level of the speaker from the audios, we asked participants to do the following:

Q1. Please transcribe this audio file by writing down what the speaker is saying. We asked the participants to transcribe only the first audio randomly presented to them. This was an open-form question. Participants were asked to write only the words they recognized if they couldn't recognize some of the words. If they couldn't recognize any words, we advised them to write "could not understand."

Q2. Please rate how confident the speaker sounds, on a 5-point Likert scale from "Not at all" to "Very much," adapted from Jiang and Pell [53, 55]. We also included the option "I don't know," in case participants could not determine the confidence level.

3.3 Recruitment

We hosted our surveys on Qualtrics (an accessible survey platform) for two months. We recruited our sighted participants from Prolific for our pilot and main studies because it provided easy access to a large pool of people [79]. Respondents were required to be 1) residents of the United States; 2) 18 years or older; 3) "workers" of Prolific with an approval rating of at least 95%; and 4) without any hearing disabilities. Respondents were asked to use headphones and to perform the study in a quiet environment. They also were asked to confirm that they had normal hearing and were native or bilingual, or professional-level English speakers. To ensure high-quality data, we added a Captcha at the beginning of the survey to discourage bots from the Prolific responses. Additionally, we

[1] To generate the sped-up stimuli, we used the website mp3cut.net: https://mp3cut.net/change-speed

included two audio attention-check questions in the survey to detect inattentive respondents [16] (e.g., "Please select *a lot* from the scale below"). After removing the responses from participants who provided wrong answers for one or more attention checks, we were left with 170 responses (out of a total of 191) that were used for further analysis. Of these, 43 sighted participants received the default speedup, 41 received the 1.5x speedup, 44 received the 2x speedup, and 42 received the 2.5x speedup condition.

For the survey with visually impaired participants, we circulated our recruitment sign-up form through email lists of various national and local organizations for the blind. We also applied snowball sampling by asking our participants to share our study invitations with others. Participants were asked in the recruitment form to sign up only if they met the following criteria: participants had to be 1) living in the United States; 2) 18 years of age or older; 3) identified as visually impaired; 4) screen reader users; 5) without hearing loss; and 6) native or bilingual, professional-level English speakers. Participants who responded through the sign-up form were screened by the researchers and we emailed each qualified participant a unique survey link. Each participant could complete the survey only once because the link was not reusable. After removing the incomplete responses, our final sample for the study was composed of 151 participants with visual impairments (out of 186 responses). Of these, 39 participants received the default speedup, 36 received the 1.5x speedup, 40 received the 2x speedup, and 36 received the 2.5x speedup condition.

Participant demographics for the two surveys are listed in Table 1. Among the visually impaired participants, 97 (64.2%) were totally blind, whereas 54 (35.8%) lived with different levels of visual impairments, such as "No vision in right eye, less than 10% in left eye." The majority of participants, (105, 69.5%) had been visually impaired since birth, whereas the rest became visually impaired afterward: 23 (15.2%) since childhood, 12 (7.9%) since early adulthood (18–40 years old), 9 (6.0%) since middle adulthood (41–60 years old), and 2 (1.3%) since late adulthood (61+ years old). More than half (86, 57.3%) of the visually impaired participants were advanced-level screen reader users, 61 (40.7%) were intermediate-level users, and only 3 (2.0%) were beginner-level users. The majority (124, 82.1%) of our participants reported that they exclusively relied on screen reader audio.

3.4 Compensation and Ethical Considerations

Regardless of whether we used their responses, each Prolific participant was paid $2.50 to complete the 10–15 min survey. The payment amount is in line with – and surpasses – the recommendation in Silberman et al. [99] to pay workers at least minimum wage in the study's location. Through a pilot study, we also confirmed this was fair compensation according to the participants (through a multiple-choice question specifically asking about fair compensation) based on the amount of work.

For the survey with visually impaired participants, each participant was paid $6 in Amazon e-gift certificates as their average completion time was 30-35 mins. We emailed them the link to the e-gift certificates within three to five days of completing the survey. Our protocol was approved by our institution's ethics review board.

	Visually impaired	Sighted
Gender		
Female	92 (60.9%)	84 (49.7%)
Male	50 (33.1%)	83 (49.1 %)
Non-binary	–	8 (2.8 %)
Age		
18-29	16 (10.6%)	79 (46.8%)
30-49	57 (37.8%)	71 (42%)
50-older	78 (51.6%)	19 (11.2%)
Education		
High school	14 (9.3%)	22 (13%)
Some college	39 (25.8%)	57 (33.7%)
Bachelors	52 (34.4%)	70 (41.4%)
Masters	37 (24.5%)	11 (6.5%)
Doctorate	4 (2.6%)	2 (1.2%)

Table 1: Demographic information of participants.

3.5 Pilot study

Prior to the initiation of our main study, we performed a pilot study with N = 20 sighted respondents from Prolific for each speedup condition. The initial analysis of the pilot showed a high correlation between the ratings of audios that conveyed confidence and neutral tone. Hence, we excluded the neutral audios from our final study, resulting in 36 audios conveying two confidence levels (confidence and doubt). The final stimuli set contained 144 audio recordings (36 audios from each playback speedup).

3.6 Accessibility

To ensure the accessibility of the survey to the participants with visual impairments, we conducted online Zoom interviews with three visually impaired participants (one female and two males). During the interview session, they were asked to respond to the survey in the presence of one of the researchers. Then a follow-up semi-structured interview session was performed by the researcher to identify any accessibility issues faced by the participants. Participants used Jaws and NVDA as screen readers and Mozilla Firefox, Google Chrome, and Microsoft Edge as browsers. The pilot study took around 40–60 minutes for each participant. Participants were each compensated with a $20 Amazon e-gift card for taking part in the pilot. As a result of feedback received during the pilot, we replaced the audio player interface with a simple button to improve accessibility. We also added a two-second pause at the beginning of each recording to avoid any overlap with the screen reader.

3.7 Procedure

During the actual experiment, participants first agreed to our consent form, then answered four screening questions. Next, we gave instructions on how to respond to the survey questions with sample audio. After the instructions, we asked them to transcribe only one audio randomly presented from the datasetWe then presented the 36 audios from one of the speedup groups (the speed condition was conducted between subjects) in random order (the confidence and doubt condition was conducted within subjects) and asked the participants to rate how confident the speaker sounded. Finally, we asked five demographic questions, how frequently they listened to or watched sped-up audio or video, and their preferred speedup rate while listening to them. To sighted participants, we showed a random code to enter into their Prolific account to receive remuneration. To the participants with visual impairments, we asked additional questions about their level of visual impairment; which assistive technologies, smart voice assistants, and screen readers they used and how frequently; and their screen reader proficiency.

3.8 Data analysis

For quantitative analysis, our data failed to meet the assumptions of normality and equal variance of errors in parametric tests. Hence, we used non-parametric versions for our statistical tests. We had one dependent variable (level of confidence ratings of the audio recordings) and one between-subject independent variable (playback speed). To observe the effects of different factors on the perceived confidence and doubt ratings, we conducted our analyses using linear mixed-effects models [75] with fixed slopes and random intercepts for each participant and each audio file. We conducted a series of multiple mixed-effect models and compared them using the Akaike information criterion (AIC) [96] to determine which factors to include. The factors explored were: playback speed, intended confidence levels (confidence or doubt), preferred speedups, screen reader proficiency, age of the participants, gender of the participants, duration of the visual impairments, gender of the speakers, and interaction terms involving them. We selected a subset of these factors that represented the best-fitting model based on the AIC values (Tables 2 and 3). We used estimated marginal means to compute the pairwise comparisons, which helped determine the significant effects across the interaction effects. The p-values were adjusted with the Tukey method [106]. Moreover, to make the doubt ratings consistent with the confidence ratings (e.g., 1 = not at all and 5 = very much), we reverse-coded the ratings for audio files conveying doubt.

4 FINDINGS

4.1 Accuracy of the audio transcriptions

We manually coded the transcription data of both PVIs and sighted participants and grouped them into four categories: 1) Correct: the transcription is fully correct; 2) Partial: participants could guess three or more words correctly; 3) Wrong: participants guessed fewer than three words correctly; 4) Could not understand: participants stated that they could not understand what the speaker was saying. For correct responses, sighted participants performed better than PVIs at default (74.42% vs 56.41%) and 2x (43.18% vs 42.50%) speed. However, for speedups 1.5x and 2.5x, PVIs transcribed the audios slightly more accurately compared to the sighted participants (1.5x: 55.56% vs 51.22%; 2.5x: 17.14% vs 14.29%). For default and 1.5x speedups, PVIs responded "could not understand" slightly more frequently compared to the sighted participants (default: 5.13% vs 4.65; 1.5x: 8.33% vs 7.32%). However, for 2x and 2.5x speeds, PVIs responded "could not understand" less frequently than sighted participants (2x: 25.00% vs 29.55%; 2.5x: 37.14% vs 45.24%).

4.2 Preferences of playback speeds

Among the participants, 108 (63.3%) sighted participants and 120 (79.5%) visually impaired participants reported using some levels of speedups while listening to audio or video content. The preferred speeds by the sighted participants include 1x $(62, 36.7\%)$, 1.25x $(47, 27.8\%)$, 1.5x $(31, 18.3\%)$, and 2x $(7, 4.1\%)$. The preferred speeds by the visually impaired participants include 0.25x $(12, 8.1\%)$, 0.5x $(13, 8.7\%)$, 1x $(16, 10.7\%)$, 1.25x $(22, 14.8\%)$, 1.5x $(38, 25.5\%)$, and 2x $(17, 11.4\%)$. Only 4 (2.6%) participants with visual impairments reported preferring speedups higher than 2x. The majority of the sighted participants preferred the default speed level and their second most preferred speedup is 1.25x while listening to audio or video content. Most of the PVIs usually preferred 1.5x speedup and 1.25x was the second most preferred speedup.

4.3 Perception of confidence and doubt at different playback speedups

The omnibus test involving the mixed effect model for participants with visual impairments is shown in Table 2. Table 3 represents the omnibus test involving the mixed effect model for sighted participants. This section presents how the perception of intended confidence levels differs based on playback speeds.

4.3.1 *Participants with visual impairments.*

Our results show that playback speed significantly predicted the perceived confidence level ($F(3, 146) = 3.01, p < 0.05$), but this effect is qualified by a higher-order interaction effect involving intended confidence levels ($F(3, 4942) = 9.10, p < 0.0001$). This finding indicates that the perceived confidence ratings differ for audios conveying doubt and confidence at different playback speeds. To observe how accurately participants with visual impairments perceived the intended confidence level of the speakers at different playback speeds, we conducted post hoc pairwise tests. Pairwise tests showed that for audio recordings conveying confidence, PVIs perceived the confidence of the speakers significantly more accurately for 1.5x speed compared to the default ($t = 3.055, d = 0.39^2, p < 0.05$) and 2.5x ($t = 3.134, d = 0.37, p < 0.05$). All other comparisons were non-significant (all $p > 0.05$).

For audio recordings conveying doubt, the pairwise tests showed that PVIs perceived the intended doubt from the audios significantly less accurately at 2x speed compared to default ($t = 3.104, d = 0.46, p < 0.01$) and 1.5x speeds ($t = 2.625, d = 0.39, p < 0.05$). All other comparisons were non-significant (all $p > 0.05$). The details are reported in Table 4. Figure 1 demonstrates the differences in perceived confidence levels across different playback speedups.

4.3.2 *Sighted participants.*

We observed that playback speed significantly predicted the perceived confidence ratings of the sighted participants ($F(3, 170) = 7.73, p < 0.01$). To observe how closely the sighted participants perceived the intended confidence at different playback speeds, we conducted pairwise comparisons. We found that sighted participants perceived the intended confidence significantly less accurately at 2.5x speed than the default ($t = 5.085, d = 0.5, p < 0.0001$), 1.5x ($t = 3.110, d = 0.31, p < 0.05$), and 2x ($t = 3.724, d = 0.36, p < 0.005$) speeds. There were no statistically significant differences in perceived confidence ratings between the pairwise comparisons of default, 1.5x, and 2x speedups (all $p > 0.05$).

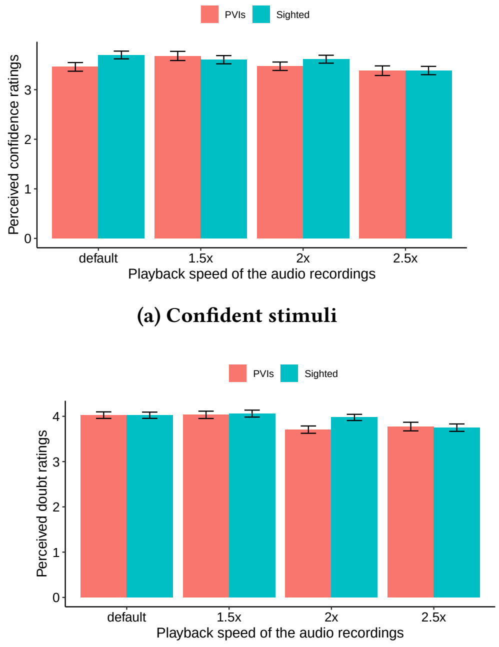
[2] Standardized effect size Cohen's d: 0.2=small effect, 0.5=medium effect, and 0.8=large effect.

	Sum Sq	Mean Sq	DoF	DenDoF	F statistic	η_p^2
Confidence level	3.55	3.55	1.00	37.89	4.03	0.10
Playback speed	7.97	2.66	3.00	146.04	3.01^{*}	0.06
Screenreader proficiency	0.72	0.72	1.00	146.70	0.81	<0.01
Preferred speedup	7.92	2.64	3.00	145.91	2.99^{*}	0.06
Confidence level : Playback speed	24.11	8.04	3.00	4942.18	9.10^{***}	<0.01
Confidence level : Screenreader proficiency	17.21	17.21	1.00	4943.35	19.49^{***}	<0.01
Confidence level : Preferred speedup	40.96	13.65	3.00	4942.37	15.47^{***}	<0.01
Confidence level : Playback speed : Screenreader proficiency	1.82	0.61	3.00	4943.37	0.69	<0.01
Confidence level : Playback speed : Preferred speedup	27.98	3.11	9.00	4942.59	3.52^{***}	<0.01

Table 2: Type III ANOVA Table (with Satterthwaite's method). ($* = p < 0.05, ** = p < 0.01, *** = p < 0.001$). The effect size η_p^2 (partial η^2) can be interpreted as small if $\eta_p^2 = 0.01$, medium if $\eta_p^2 = 0.06$, and large if $\eta_p^2 = 0.14$ [62] (PVIs) .

	Sum Sq	Mean Sq	DoF	DenDoF	F statistic	η_p^2
Confidence level	8.06	8.06	1.00	37.09	9.00^{**}	0.20
Playback speed	20.76	6.92	3.00	170.24	7.73^{***}	0.12
Preferred speedups	9.03	4.52	2.00	169.96	5.04^{**}	0.06
Confidence level : Playback speed	5.44	1.81	3.00	5759.65	2.02	<0.01
Confidence level : Preferred speedups	16.29	8.14	2.00	5759.39	9.09^{***}	<0.01
Confidence level : Playback speed : Preferred speedups	18.99	3.16	6.00	5758.54	3.53^{**}	<0.01

Table 3: Type III ANOVA Table (with Satterthwaite's method). ($* = p < 0.05, ** = p < 0.01, *** = p < 0.001$). The effect size η_p^2 (partial η^2) can be interpreted as small if $\eta_p^2 = 0.01$, medium if $\eta_p^2 = 0.06$, and large if $\eta_p^2 = 0.14$ (Sighted).



(a) Confident stimuli

(b) Doubt stimuli

Figure 1: Mean confidence ratings with 95% CI across different playback speeds and intended confidence levels of the speakers (PVIs and Sighted).

Participants	Audio types	Speed	Mean	Std.	95%CI
PVIs	Confident audios	default	3.46	1.172	0.088
		1.5x	3.681	1.184	0.092
		2x	3.473	1.138	0.085
		2.5x	3.383	1.208	0.097
	Doubt audios	default	1.974	0.974	0.073
		1.5x	1.967	1.039	0.081
		2x	2.294	1.082	0.081
		2.5x	2.226	1.192	0.096
Sighted	Confident audios	default	3.7	1.094	0.078
		1.5x	3.605	1.13	0.083
		2x	3.615	1.136	0.08
		2.5x	3.387	1.159	0.084
	Doubt audios	default	1.976	0.967	0.069
		1.5x	1.939	1.043	0.077
		2x	2.024	0.979	0.069
		2.5x	2.25	1.128	0.082

Table 4: Mean, standard deviation, and 95% confidence intervals for different speeds across audio types (PVIs and Sighted).

Figure 1 shows the perceived confidence ratings by the sighted participants across each playback speed.

Similarly, for audio recordings conveying doubt, the pairwise tests showed that sighted participants perceived the intended doubt significantly less accurately at 2.5x speeds compared to the default ($t = 3.040, d = 0.3, p < 0.05$), 1.5x ($t = 2.941, d = 0.3, p < 0.05$), and 2x ($t = 2.768, d = 0.27, p < 0.05$) speeds. All other comparisons were non-significant (all $p > 0.05$). The details are reported in Table 4.

4.3.3 Differences in perception of confidence and doubt by PVIs and sighted participants. Next, we explored how the perception of confidence and doubt differ by visually impaired and sighted participants at different playback speeds. The differences between the two groups are shown in Figure 1. The overall results revealed that overall sighted participants perceived the intended confidence ($t = 2.998, d = 0.09, p < 0.005$) and doubt ($t = 2.450, d = 0.07, p < 0.05$) significantly more accurately than PVIs.

For audios conveying confidence, our pairwise comparisons showed that sighted participants perceived the intended confidence more accurately at default ($t = 3.518, d = 0.23, p < 0.001$), and 2x speedups ($t = 4.065, d = 0.27, P < 0.0001$). However, PVIs perceived the intended confidence more accurately at 1.5x ($t = 2.163, d = 0.15, P < 0.05$) speedup compared to sighted participants. No statistically significant difference was observed between the two groups at speed 2.5x ($p > 0.05$).

For doubt-conveying audios, we found that the perception of doubt by PVIs and sighted participants significantly varied only at 2x speed ($t = 5.939, d = 0.39, p < 0.0001$). The sighted participants perceived intended doubt at 2x speedup significantly better than the PVIs did. There were no statistically significant differences observed in doubt perception by the two groups at speed default, 1.5x, and 2.5x (all $p > 0.05$).

4.4 Factors impacting the perception of confidence and doubt - visually impaired participants

4.4.1 Preferred speedups. To simplify the analysis, we categorized the participants into two groups based on their preferred playback speedups while listening to audio and video content: default (or 1x) speed preferred and accelerated speed preferred. Participants who preferred speedups from 1.25x to 2x were in the "accelerated speed preferred" group. Our findings indicated that preferred speedups of the PVIs significantly predicted the perceived confidence ratings ($F(3, 145) = 2.99, p < 0.05$). However, this effect is qualified by a higher-order interaction effect involving intended confidence levels and playback speedups ($F(9, 4942) = 3.52, p < 0.001$).

To observe the effect of preferred speedups on the perception of doubt, our findings showed that PVIs who usually preferred accelerated speedups while listening to audio and video content perceived doubt conveyed by the speakers significantly more closely ($t = 4.321, d = 0.6, p < 0.0001$) compared to the PVIs who preferred default speed. The differences in perception of doubt at different speedups by the two groups are illustrated in Figure 2.

To further observe the effect of preferred speedups of the participants across each speedup condition, our pairwise comparisons indicated that PVIs who usually preferred accelerated speedups perceived the intended doubt significantly more closely at default ($t = 3.397, d = 0.6, p < 0.001$), 1.5x ($t = 2.984, d = 0.75, p < 0.01$), and 2x ($t = 2.084, d = 0.87, p < 0.05$) speedups compared to the PVIs who usually preferred default speed. At 2.5x speed, the difference in the perception of doubt was not statistically significant between the two groups ($p > 0.05$).

To observe the effect of preferred speedups in confidence perception by the two groups for audios conveying confidence, we found no statistically significant differences ($p > 0.05$).

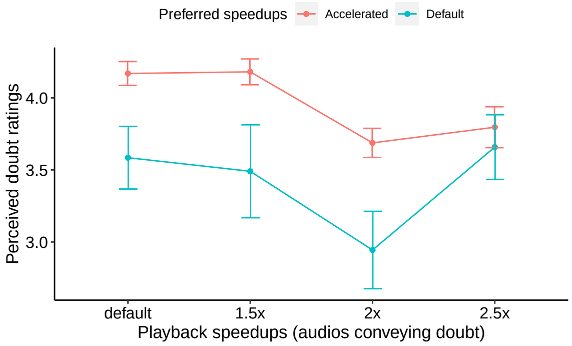


Figure 2: Mean confidence ratings with 95%CI for audios conveying doubt across playback speedups and preferred speedups (PVIs).

4.4.2 Screen reader proficiency. To observe the effects of screen reader proficiency on the perception of confidence, we considered only the PVIs with advanced and intermediate levels of screen reader proficiency. We excluded the participants with beginner-level proficiency because the sample size was very small (only three participants) compared to the other two groups. As reported in Table 2, the interaction effect of intended confidence levels and screen reader proficiency of the PVIs significantly predicted the perceived confidence ratings ($F(1, 4943) = 19.49, p < 0.0001$).

For audios conveying confidence, we observed statistically significant differences in confidence perception by the advanced screen reader users compared to the intermediate screen reader users ($t = 2.644, d = 0.19, p < 0.005$). Participants with advanced screen reader proficiency rated the audio recordings conveying confidence more accurately than the participants with intermediate screen reader proficiency. The pairwise comparisons across each playback speedup showed that PVIs who were advanced screen reader users perceived the intended confidence significantly more accurately at default ($t = 1.973, d = 0.3, p < 0.05$), and 1.5x ($t = 2.538, d = 0.38, p < 0.05$) speedups compared to the PVIs with intermediate screen reader proficiency. All other comparisons were non-significant (all $p > 0.05$).

For audios conveying doubt, we did not observe any significant differences in perception of doubt between the PVIs with advanced and intermediate levels of screen reader proficiency ($p > 0.05$).

4.5 Factors impacting the perception of confidence and doubt – sighted participants

4.5.1 Preferred speedups. We similarly categorized the sighted participants into two groups based on their preferred playback speedups while listening to audio and video content: default and accelerated speeds preferred. Our results showed that preferred speedups of the sighted participants significantly predicted the perceived confidence ratings ($F(2, 170) = 5.04, p < 0.01$). However, this effect is qualified by a higher-order interaction effect involving intended confidence levels and playback speedups ($F(6, 5758) = 3.53, p < 0.01$).

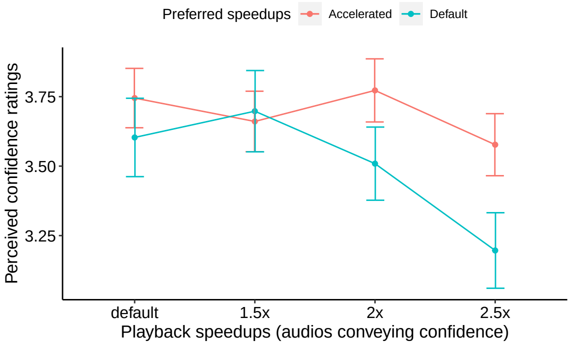


Figure 3: Mean confidence ratings with 95% CI for audios conveying confidence across playback speeds and preferred speedups (Sighted).

To observe the differences in confidence perception by the two groups for audios conveying confidence, we observed statistically significant differences ($t = 3.432, d = 0.2, p < 0.001$). To further observe the effect of preferred speedups on confidence perception across different playback speedups, our pairwise comparisons indicated that sighted participants who usually preferred accelerated speedups perceived the intended confidence significantly more accurately at the playback speedups 2x ($t = 2.530, d = 0.28, p < 0.05$) and 2.5x ($t = 3.720, d = 0.42, p < 0.001$) compared to the participants who preferred the default speed. There were no statistically significant differences between those who preferred default and 1.5x speeds ($p > 0.05$). The differences in confidence perception for different preferred speedups across different playback speedups are shown in Figure 3. We found no statistically significant differences ($p > 0.05$) between the sighted participants who usually preferred accelerated speedups while listening to audio and video content compared to the participants who preferred default speedup.

5 DISCUSSION

We first summarize our key findings and later discuss the broader implications of the findings.

5.1 Key findings

Our results show how people with visual impairments and sighted people perceive confidence and doubt in the voice of speakers in accelerated speech and which factors influence their perceptions. Our sighted participants were able to perceive the confidence and doubt conveyed by the speakers almost equally well up to a 2x speedup. Their performance significantly decreased at 2.5x speed. However, the perception of confidence and doubt by participants with visual impairments at different playback speeds was not straightforward. PVIs were able to perceive the confidence and doubt conveyed by the speakers comparatively more accurately up to a 1.5x speedup. While perceiving the intended confidence, surprisingly, they performed better at 1.5x speed than the default speed. Their performance decreased significantly at a 2x speedup when perceiving doubt.

Prior works also explored the differences in voice-based perception by blind and sighted people. For example, Martins et al. [70] observed that sighted and blind people performed equally well at identifying an interrogative or declarative tone of voice. Additionally, sighted people performed significantly better compared to blind people when identifying emotionally incongruent items (e.g., happy content said in sad voices). Oleszkiewicz et al. [76] reported that blind people process socially relevant information (such as competence and warmth) similarly to sighted people from nonverbal voice cues. Consistent with prior research, our study revealed that both visually impaired and sighted participants performed similarly when perceiving doubt at various speedups. Overall, sighted participants performed marginally better than individuals with visual impairments at a 2x speedup in perceiving confidence and doubt. However, participants with visual impairments demonstrated superior performance in perceiving confidence at a 1.5x speedup in comparison to their sighted counterparts.

Prior research has suggested that the listening rate of individuals with visual impairments may be influenced by their use of screen readers [28, 71]. Specifically, individuals who have had early exposure and more experience with screen readers may exhibit better listening rates compared to others [28, 71]. Consistent with this literature, our study found that participants with advanced screen reader proficiency displayed more accurate perceptions of speaker confidence compared to intermediate screen reader users. Furthermore, we discovered a relationship between preferred speedups and the perception of doubt among visually impaired participants. Specifically, participants who typically preferred accelerated speeds when consuming audio and video in their daily lives were more likely to perceive doubt in the voices of speakers more accurately at higher speeds, in contrast to those who preferred the default speed.

5.2 Implications

5.2.1 *Appropriate speed to convey expressive content.*

Individuals with visual impairments tend to interact with audio-based input and output at a faster speed, and sighted individuals also often prefer higher speeds while consuming audio visual content [38, 73]. Our findings have implications for selecting an appropriate speedup when interacting with audio or videos that contain expressive content, such as confidence and doubt. As we found, listeners may not be able to accurately perceive such expressive content at higher speedups (such as 2–2.5x). Our findings indicate that PVIs performed similarly or slightly better than sighted individuals in transcribing audio, but this may not necessarily apply in the context of conveying emotional cues through voices in accelerated speech. Previous works have emphasized incorporating faster speech in conversational agents to accommodate the accessibility needs of individuals with visual impairments [17, 28, 32, 90], or adjusting the playback speed adaptively while consuming video content [31, 61]. These suggestions included adjusting the playback speed based on the complexity of the content or areas of interest. Nonetheless, *we emphasize that future technologies should consider incorporating appropriate speedup rates and improving techniques for accelerating audio in a manner that conveys emotional features, such as confidence levels, using audio prosodic cues, such as tone and pitch, so that the cues can be understood at accelerated speeds.*

5.2.2 Humanizing audio interaction in synthesized accelerated speech. As synthetic speech becomes more ubiquitous, recent research has been dedicated to making it more humanized and natural [39, 42, 67, 98]. Many visually impaired individuals rely heavily on synthetic speech to consume content, much of which contains emotional expression. For instance, a significant portion (35%) of the text in Massive Open Online Courses (MOOCs) contains emotional expression [52]. Prior research has also demonstrated that human-voice audio description (AD) is better at conveying sadness than text-to-speech (TTS) AD [40, 41]. Similarly, in the context of voice-based smart assistants, Cohn et al. [33] found that a human voice is perceived as better at conveying emotional valence, such as happiness, than the voice used by Alexa. However, as people with visual impairments heavily rely on synthetic speech at faster speeds, future research should aim to make synthetic speech more human-like and expressive in accelerated speech. Prior work has revealed that individuals with visual impairments perceive the voice of conversational agents as more machine-like in accelerated speech [32]. Our findings indicate that both PVIs and sighted individuals perceive confidence levels less accurately as playback speeds increase. Therefore, we argue that *future research should aim to understand how emotional features, such as confidence and doubt, can be better conveyed in accelerated synthesized speech and how to make them more human-like and expressive at faster speeds.* Furthermore, our exploration of confidence perception from human recorded voices can be used to design more human-like synthesized speech.

5.2.3 Effectively communicate AI-based decisions in assistive technologies. Prior studies found that PVIs are skeptical about the accuracy of the automatic assessments of AI-based assistive technologies and have limited trust in them [20, 21]. To gain users' trust in AI-based systems, prior works suggested making the factors that influenced the decision-making of the algorithm visible, transparent, and easily understandable for users [34, 58, 101]. Conveying the probability of algorithmic predictions can increase the transparency of and trust in a system [51]. It is important to convey algorithmic decisions as clearly as possible to convey the fallibility of these systems [21, 82], thus preventing unrealistic expectations that could lead to a negative experience. Our work also provides novel insights into the potential of effectively providing the confidence level of the algorithmic decisions incorporated in the system voice in accelerated speed to PVIs and making the systems more transparent and usable. Receiving confidence cues through speech will provide more control to PVIs and help them make better decisions on whether to trust (or not) the output of the AI-based assistive technologies. It will also help PVIs to avoid potential harm or embarrassing situations and hence increase the social acceptability of AI-based assistive systems.

5.2.4 Effect of listeners' characteristics on the perception of confidence and doubt. Prior works observed how listener characteristics influenced people's voice-based assessments and listening rates [25, 28, 77]. Our study extends this research by examining how preferred speedups and screen reader proficiency influence the perception of confidence and doubt. We observed that sometimes the characteristics of the listeners are associated with their perception of confidence and doubt differently. For example, we found that preferred speedup was associated with better doubt perception but not with the perception of confidence by PVIs. In contrast, screen reader proficiency impacted confidence perception but not the perception of doubt. We also observed differences between the two participant groups. Unlike PVIs, the preferred speedup was associated with better confidence perception for sighted participants but not for their perception of doubt. *Therefore, we argue that more research is needed to understand these nuances of how the characteristics of listeners are associated with their perception of confidence and doubt and why.*

5.2.5 Voice-based assessments by blind and sighted people. Prior works examined how the voice pitch of speakers influences listeners' assessments of various social traits (e.g., attractiveness, masculinity or femininity, and dominance) [60, 88, 91]. Oleszkiewicz et al. explored whether there are differences in the judgment of social traits such as trustworthiness, competence, and warmth by blind and sighted people [76]. They reported that both blind and sighted people rated voices with lowered pitch as being more competent and trustworthy than voices with raised pitch. Similarly, we observed how blind and sighted people assessed the confidence level of the speakers from the tone and pitch of their voices in different playback speedups. We found that both visually impaired and sighted participants considered voices with a rising pitch less confident than voices with a low pitch. Perceiving the confidence level of the speakers and other traits may be particularly important for blind people. For example, they may often rely on the opinions and assistance of other people and on algorithmic decision-based assistive technologies in everyday life because of their lack of access to visual cues [20, 21, 64]. *Hence, our findings provide directions for future research to explore the differences in voice-based assessments of different social and emotional traits (e.g., sarcasm, sincerity, happiness) by visually impaired and sighted people.*

5.2.6 Human recorded vs. synthesized speech. According to prior research, people with visual impairments typically understand faster speech better than sighted people [28, 102]. However, we observed that sighted participants were able to perceive confidence and doubt from speech reasonably up to the 2x speedup rate, and PVIs could perceive confidence and doubt up to the 1.5x rate. One possible explanation could be that PVIs are much better at understanding and have extensive interaction with synthesized speech [73, 102]. Moose and Trouvain observed that blind people understood synthesized speech better than sighted people [73]. Usually, people who became blind recently prefer voices and speeds resembling human speech (concatenative synthesis), while experienced screen reader users prefer more robotic-sounded high-speed voices (formant synthesis) [28, 71]. However, the research related to this question is still inconclusive. For example, Papadopoulos et al. observed no significant differences between sighted and visually impaired participants regarding the comprehension of natural and synthetic speech [81]. In the context of voice recognition tasks, prior works did not find any significant differences in the performance of blind and sighted people [45, 49, 107]. *Therefore, more research is needed to understand how people perceive confidence levels from voices in synthesized vs. natural speech.*

6 LIMITATIONS AND FUTURE WORK

We recognize several limitations of our study that could be addressed in future work. Our visually impaired participant sample was small, limited to recruits from a few national foundations for the blind, and restricted to those who chose to respond to our ad, so it is difficult to know how well our findings generalize to the greater population. However, we also note the challenges in reaching this population and, compared to other recent studies of privacy concerns for the visually impaired, our sample size is relatively large [18, 21, 28]. Moreover, our findings related to confidence perception by people with visual impairments are not universal and may not be generalized to the perception of confidence and doubt by people with other accessibility needs. For example, people with autism may perceive confidence and other emotional cues from speech differently [44, 68]. Future work should explore how people with different accessibility needs perform in perceiving confidence and emotions in accelerated speech. We acknowledge the potential limitations of our study, particularly in the recruitment of our two participant groups. It should be noted that the groups were recruited through different methods, with one group sourced from paid crowdsourcing platforms and the other from trusted community organizations. Additionally, differences in compensation and participation time were also present. Nevertheless, it is important to note that the practice of recruiting participants from diverse platforms has been utilized and compared in the accessibility community [21, 47, 87].

In this study, we considered human-recorded audio instead of synthesized audio recordings. As PVIs have a long history and experience with synthesized speech, future work should explore the perception of speaker confidence in synthesized speech. Our stimuli contained only three sentences with two confidence levels recorded by six actors. Our intention in this research was to capture participants' perception of the confidence level of the speakers from the tone of the voices rather than measuring the intelligibility or listening rates of the participants. However, future work should explore a larger dataset with more variations. For example, it would be interesting to investigate how sentences with lexical cues (such as "definitely," "maybe," or "perhaps") affect the perception of confidence along with the tone of the voice. Finally, we explored people's perception of only the confidence level of the speakers (confidence and doubt) through voice-based cues. It would be interesting to study the perception of other emotional cues (such as happiness and sadness) through voice-based assessments.

7 CONCLUSIONS

AI-based assistive technologies for people with visual impairments (PVIs) can convey inaccurate information and misrepresent people while providing assistance. This issue is particularly relevant for marginalized communities, such as Black, Indigenous, Persons of Color (BIPOC), non-binary, and transgender people, who have expressed concerns about being misrepresented by AI-generated image descriptors. In order to mitigate this issue, previous research has suggested framing image descriptions in a negative light to increase user awareness of the potential fallibility of automatically generated image captions. As PVIs interact with assistive systems primarily through audio, conveying the different levels of confidence through the intonation and wording of the output sentence can help make users aware of the fallibility of AI-based assistive technologies. Additionally, PVIs are accustomed to using accelerated speech and receiving feedback through short sentences from assistive systems. Therefore, investigating how effectively confidence or doubt can be conveyed through short and accelerated speech, and how accurately people can perceive these cues, is crucial.

We conducted online surveys with PVIs (n =151) and sighted participants (n =170) to explore their perceptions of a speaker's confidence or doubt in accelerated speech. We examined how the accuracy of their perceptions varied based on different playback speedups (default to 2.5x). Our findings suggest that while visually impaired participants could perceive the confidence levels of the speakers more closely up to a 1.5x speedup rate, sighted participants were able to perceive the confidence levels in speech up to a 2x speedup. We also observed that the preferred speedup is associated with the confidence and doubt perception of the participants. For example, PVIs who preferred accelerated speedups while listening to audio or video content in their daily lives perceived doubt conveyed by the speakers more accurately compared to the PVIs who preferred default speeds.

Our findings have implications for the careful design of future audio output-based assistive technologies and conversational agents. These agents should incorporate a confidence feature to convey the level of confidence of their underlying algorithm to make PVIs aware of its possible fallibility and help them make better decisions. In particular, future systems should consider appropriate speedup rates to convey confidence and doubt (e.g., 1.5–2x speedups) instead of the speedup PVIs prefer while interacting with screen readers. In the longer term, research is needed to improve the preservation of speech characteristics in accelerated speech, so that expressive contents (e.g., confidence and doubt) can be conveyed at fast speech rates. Overall, the findings of this study highlight the importance of considering the nuances of human perception in designing assistive technologies for PVIs.

REFERENCES

[1] 2016. Fast-Forwarding is Becoming a New Way to Watch Videos. Last accessed April 2024, https://thenewstack.io/fast-forwarded-videos-wave-future/.

[2] 2018. Average Speaking Rate and Words per Minute. Last accessed August 2022, https://virtualspeech.com/blog/average-speaking-rate-words-per-minute.

[3] 2022. Aipoly. Last accessed August 2022, www.aipoly.com.

[4] 2022. Amazon Echo. Last accessed August 2022, https://www.amazon.com/echo/s?k=echo.

[5] 2022. Apple Voiceover: Hear what's happening on your screen. Last accessed August 2022, https://www.apple.com/accessibility/vision/.

[6] 2022. Clear Reader Plus. Last accessed August 2022, https://enablingtechnology.com/text-readers/clear-reader.

[7] 2022. Freedom Scientifc: JAWS. Last accessed August 2022, https://www.freedomscientific.com/products/software/jaws/.

[8] 2022. Google Home. Last accessed August 2022, https://home.google.com/welcome/.

[9] 2022. KNFB Reader. Last accessed August 2022, https://nfb.org/programs-services/knfb-reader.

[10] 2022. NV Access. Last accessed August 2022, https://www.nvaccess.org/.

[11] 2022. Orcam. Last accessed August 2022, www.orcam.com/en/.

[12] 2022. Seeing AI. Last accessed August 2022, www.microsoft.com/en-us/seeing-ai.

[13] 2022. TalkBack. Last accessed August 2022, https://play.google.com/store/apps/details?id=com.google.android.marvin.talkback&hl=en..

[14] 2022. YouTube Says That an Increasing Number of Viewers are Watching Videos at Faster Playback Speeds. Last accessed April 2024, https://www.socialmediatoday.com/news/youtube-says-that-an-increasing-number-of-viewers-are-watching-videos-at-fa/630657.

[15] 2023. Introducing ChatGPT. Last accessed April 2024, https://openai.com/blog/chatgpt.

[16] James D Abbey and Margaret G Meloy. 2017. Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management* 53 (2017), 63–70.

[17] Ali Abdolrahmani, Ravi Kuber, and Stacy M. Branham. 2018. "Siri Talks at You": An Empirical Investigation of Voice-Activated Personal Assistant (VAPA) Usage by Individuals Who Are Blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2018, Galway, Ireland, October 22-24, 2018*. ACM, 249–258. https://doi.org/10.1145/3234695.3236344

[18] Tousif Ahmed, Apu Kapadia, Venkatesh Potluri, and Manohar Swaminathan. 2018. Up to a limit? privacy concerns of bystanders and their willingness to share additional information with visually impaired users of assistive technologies. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–27.

[19] Taslima Akter. 2020. Privacy Considerations of the Visually Impaired with Camera Based Assistive Tools. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 69–74.

[20] Taslima Akter, Tousif Ahmed, Apu Kapadia, and Manohar Swaminathan. 2022. Shared Privacy Concerns of the Visually Impaired and Sighted Bystanders with Camera-Based Assistive Technologies. *ACM Transactions on Accessible Computing (TACCESS)* 15, 2 (2022), 1–33.

[21] Taslima Akter, Tousif Ahmed, Apu Kapadia, and Swami Manohar Swaminathan. 2020. Privacy Considerations of the Visually Impaired with Camera Based Assistive Technologies: Misrepresentation, Impropriety, and Fairness. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–14.

[22] Chieko Asakawa, Hironobu Takagi, Shuichi Ino, and Tohru Ifukube. 2003. Maximum listening speeds for the blind. (01 2003).

[23] Vikas Ashok, Yevgen Borodin, Svetlana Stoyanchev, Yury Puzis, and I. V. Ramakrishnan. 2014. Wizard-of-Oz evaluation of speech-driven web browsing interface for people with vision impairments. In *International Web for All Conference, W4A '14, Seoul, Republic of Korea, April 7-9, 2014*. ACM, 12:1–12:9. https://doi.org/10.1145/2596695.2596699

[24] Shiri Azenkot and Nicole B. Lee. 2013. Exploring the Use of Speech Input by Blind People on Mobile Devices. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (Bellevue, Washington) *(ASSETS '13)*. Association for Computing Machinery, New York, NY, USA, Article 11, 8 pages. https://doi.org/10.1145/2513383.2513440

[25] Marialena Barouti, Konstantinos Papadopoulos, and Georgios Kouroupetroglou. 2013. *Synthetic and natural speech intelligibility in individuals with visual impairments: effects of experience and presentation rate*. Vol. 33. 695–701. https://doi.org/10.3233/978-1-61499-304-9-695

[26] Cynthia L Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P Bigham, Anhong Guo, and Alexandra To. 2021. "It's Complicated": Negotiating Accessibility and (Mis) Representation in Image Descriptions of Race, Gender, and Disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 1–19.

[27] Yevgen Borodin, Jeffrey P Bigham, Glenn Dausch, and IV 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 (W4A)*. 1–10.

[28] Danielle Bragg, Cynthia L. Bennett, Katharina Reinecke, and Richard E. Ladner. 2018. A Large Inclusive Study of Human Listening Rates. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. ACM, 444. https://doi.org/10.1145/3173574.3174018

[29] Stacy M. Branham and Antony Rishin Mukkath Roy. 2019. Reading Between the Guidelines: How Commercial Voice Assistant Guidelines Hinder Accessibility for Blind Users. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2019, Pittsburgh, PA, USA, October 28-30, 2019*. ACM, 446–458. https://doi.org/10.1145/3308561.3353797

[30] Susan E Brennan and Maurice Williams. 1995. The feeling of another s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language* 34, 3 (1995), 383–398.

[31] Kai-Yin Cheng, Sheng-Jie Luo, Bing-Yu Chen, and Hao-Hua Chu. 2009. Smartplayer: user-centric video fast-forwarding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 789–798.

[32] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. 2020. "Nobody Speaks that Fast!" An Empirical Study of Speech Rate in Conversational Agents for People with Vision Impairments. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*. ACM, 1–13. https://doi.org/10.1145/3313831.3376569

[33] Michelle Cohn, Eran Raveh, Kristin Predeck, Iona Gessinger, Bernd Möbius, and Georgia Zellou. 2020. Differences in gradient emotion perception: Human vs. Alexa voices. In *Proceedings of Interspeech*.

[34] Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic transparency in the news media. *Digital journalism* 5, 7 (2017), 809–828.

[35] Christel Dijkstra, Emiel Krahmer, and Marc Swerts. 2006. Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence. In *Proceedings of the Speech Prosody Conference*. Citeseer.

[36] Derek Doran, Sarah Schulz, and Tarek R Besold. 2017. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794* (2017).

[37] Jeroen Dral, Dirk Heylen, et al. 2011. Detecting uncertainty in spoken dialogues: an exploratory research for the automatic detection of speaker uncertainty by using prosodic markers. In *Affective Computing and Sentiment Analysis*. Springer, 67–77.

[38] Songshuang Duan and Xiaoqian Chen. 2019. Why College Students Watch Streaming Drama at Higher Playback Speed: The Uses and Gratifications Perspective. In *2019 International Joint Conference on Information, Media and Engineering (IJCIME)*. IEEE, 401–403.

[39] Bryan Duggan and Mark Deegan. 2003. Considerations in the usage of text to speech (TTS) in the creation of natural sounding voice enabled web systems. *ISICT* 3 (2003), 433–438.

[40] Anna Fernández-Torné and Anna Matamala. 2015. Text-to-speech vs. human voiced audio descriptions: a reception study in films dubbed into Catalan. *The Journal of Specialised Translation* 24 (2015), 61–88.

[41] Louise Fryer and Jonathan Freeman. 2014. Can you feel what I'm saying? The impact of verbal information on emotion elicitation and presence in people with a visual impairment. *Proceedings of the international society for presence research* (2014), 99–107.

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

[43] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. 2018. Explainable AI: the new 42?. In *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2*. Springer, 295–303.

[44] Ofer Golan, Simon Baron-Cohen, Jacqueline J Hill, and MD17072749 Rutherford. 2007. The 'Reading the Mind in the Voice'test-revised: a study of complex emotion recognition in adults with and without autism spectrum conditions. *Journal of autism and developmental disorders* 37, 6 (2007), 1096–1106.

[45] Frédéric Gougoux, Pascal Belin, Patrice Voss, Franco Lepore, Maryse Lassonde, and Robert J Zatorre. 2009. Voice perception in blind persons: a functional magnetic resonance imaging study. *Neuropsychologia* 47, 13 (2009), 2967–2974.

[46] Frédéric Gougoux, F. Lepore, M. Lassonde, P. Voss, R. Zatorre, and P. Belin. 2004. Neuropsychology: Pitch discrimination in the early blind. *Nature* 430 (2004), 309–309.

[47] João Guerreiro and Daniel Gonçalves. 2016. Scanning for digital content: How blind and sighted people perceive concurrent speech. *ACM Transactions on Accessible Computing (TACCESS)* 8, 1 (2016), 1–28.

[48] 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 (ASSETS '15)*. Association for Computing Machinery, New York, NY, USA, 3–11. https://doi.org/10.1145/2700648.2809840

[49] Deborah Günzburger, Annebeth Bresser, and M Ter Keurs. 1987. Voice identification of prepubertal boys and girls by normally sighted and visually handicapped subjects. *Language and Speech* 30, 1 (1987), 47–58.

[50] Roy Hamilton, Alvaro Pascual-Leone, and Gottfried Schlaug. 2004. Absolute Pitch in Blind Musicians. *Neuroreport* 15 (04 2004), 803–6. https://doi.org/10.1097/00001756-200404090-00012

[51] Hendrik Heuer and Andreas Breiter. 2020. More than accuracy: towards trustworthy machine learning interfaces for object recognition. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 298–302.

[52] Garron Hillaire, Francisco Iniesto, and Bart Rienties. 2019. Humanising text-to-speech through emotional expression in online courses. *Journal of Interactive Media in Education* 1 (2019).

[53] Xiaoming Jiang, Kira Gossack-Keenan, and Marc D Pell. 2020. To believe or not to believe? How voice and accent information in speech alter listener impressions of trust. *Quarterly Journal of Experimental Psychology* 73, 1 (2020), 55–79.

[54] Xiaoming Jiang and Marc D Pell. 2014. Encoding and decoding confidence information in speech. In *Proceedings of the 7th international conference in speech prosody (social and linguistic speech prosody)*, Vol. 5762579.

[55] Xiaoming Jiang and Marc D Pell. 2017. The sound of confidence and doubt. *Speech Communication* 88 (2017), 106–126.

[56] Taenyun Kim and Hayeon Song. 2022. Communicating the Limitations of AI: The Effect of Message Framing and Ownership on Trust in Artificial Intelligence. *International Journal of Human–Computer Interaction* (2022), 1–11.

[57] Charles E Kimble and Steven D Seidel. 1991. Vocal signs of confidence. *Journal of Nonverbal Behavior* 15, 2 (1991), 99–105.

[58] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2390–2395.

[59] Emiel Krahmer and Marc Swerts. 2005. How children and adults produce and perceive uncertainty in audiovisual speech. *Language and speech* 48, 1 (2005), 29–53.

[60] Jody Kreiman and Diana Sidtis. 2011. *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

[61] Kazutaka Kurihara. 2012. CinemaGazer: a system for watching videos at very high speed. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. 108–115.

[62] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013), 863.

[63] David Lang, Guanling Chen, Kathy Mirzaei, and Andreas Paepcke. 2020. Is faster better? A study of video playback speed. In *Proceedings of the tenth international conference on learning analytics & knowledge*. 260–269.

[64] J David Lewis and Andrew Weigert. 1985. Trust as a social reality. *Social forces* 63, 4 (1985), 967–985.

[65] Francis C Li, Anoop Gupta, Elizabeth Sanocki, Li-wei He, and Yong Rui. 2000. Browsing digital video. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 169–176.

[66] Brian Y Lim and Anind K Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing*. 415–424.

[67] Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li. 2021. Expressive tts training with frame and style reconstruction loss. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1806–1818.

[68] Hope Macdonald, Michael Rutter, Patricia Howlin, Patricia Rios, Ann Le Conteur, Christopher Evered, and Susan Folstein. 1989. Recognition and expression of emotional cues by autistic and normal adults. *Journal of Child Psychology and Psychiatry* 30, 6 (1989), 865–877.

[69] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5988–5999.

[70] Ana Teresa Martins, Luís Faísca, Helena Vieira, and Gabriela Gonçalves. 2019. Emotional recognition and empathy both in deaf and blind adults. *The Journal of Deaf Studies and Deaf Education* 24, 2 (2019), 119–127.

[71] Ted McCarthy, Joyojeet Pal, and Edward Cutrell. 2013. The "Voice" Has It: Screen Reader Adoption and Switching Behavior Among Vision Impaired Persons in India. *Assistive Technology* 25, 4 (2013), 222–229. https://doi.org/10.1080/10400435.2013.768719

[72] Laura Monetta, Henry S Cheang, and Marc D Pell. 2008. Understanding speaker attitudes from prosody by adults with Parkinson's disease. *Journal of neuropsychology* 2, 2 (2008), 415–430.

[73] Anja Moos and Jürgen Trouvain. 2007. Comprehension of ultra-fast speech-blind vs." normally hearing" persons. *Proceedings of the 16th International Congress of Phonetic Sciences* (01 2007).

[74] Toru Nagahama and Yusuke Morita. 2017. Effect analysis of playback speed for lecture video including instructor images. *International Journal for Educational Media and Technology* 11, 1 (2017), 50–58.

[75] Ann L Oberg and Douglas W Mahoney. 2007. Linear mixed effects models. *Topics in biostatistics* (2007), 213–234.

[76] Anna Oleszkiewicz, Katarzyna Pisanski, Kinga Lachowicz-Tabaczek, and Agnieszka Sorokowska. 2017. Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychonomic bulletin & review* 24, 3 (2017), 856–862.

[77] Anna Oleszkiewicz, Katarzyna Pisanski, and Agnieszka Sorokowska. 2017. Does blindness influence trust? A comparative study on social trust among blind and sighted adults. *Personality and Individual Differences* 111 (2017), 238–241.

[78] Aydin Ziya Ozgur and Huseyin Selcuk Kiray. 2007. Evaluating Audio Books as Supported Course Materials in Distance Education: The Experiences of the Blind Learners. *Online Submission* 6, 4 (2007).

[79] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[80] Konstantinos Papadopoulos, Evangelia Katemidou, Athanasios Koutsoklenis, and Eirini Mouratidou. 2010. Differences among sighted individuals and individuals with visual impairments in word intelligibility presented via synthetic and natural speech. *Augmentative and Alternative Communication* 26, 4 (2010), 278–288.

[81] Konstantinos Papadopoulos and Eleni Koustriava. 2015. Comprehension of synthetic and natural speech: Differences among Sighted and visually impaired young adults. *Enabling Access for Persons with Visual Impairment* 147 (2015), 149–153.

[82] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (2019).

[83] Raymond Pastore and Albert D Ritzhaupt. 2015. Using time-compression to make multimedia learning more efficient: Current research and practice. *TechTrends* 59, 2 (2015), 66–74.

[84] Silke Paulmann, Sarah Jessen, and Sonja A Kotz. 2009. Investigating the multimodal nature of human communication: Insights from ERPs. *Journal of Psychophysiology* 23, 2 (2009), 63–76.

[85] Silke Paulmann and Marc D Pell. 2011. Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion* 35, 2 (2011), 192–201.

[86] Jonathan E Peelle and Matthew H Davis. 2012. Neural oscillations carry speech rhythm through to comprehension. *Frontiers in psychology* 3 (2012), 320.

[87] Michal Pieniak, Kinga Lachowicz-Tabaczek, Maciej Karwowski, and Anna Oleszkiewicz. 2022. Sensory compensation beliefs among blind and sighted individuals. *Scandinavian journal of psychology* 63, 1 (2022), 72–82.

[88] Katarzyna Pisanski and Gregory A Bryant. 2019. The evolution of voice perception. *The oxford handbook of voice studies* (2019), 269–300.

[89] Heather Pon-Barry and Stuart M Shieber. 2011. Recognizing uncertainty in speech. *EURASIP Journal on Advances in Signal Processing* 2011 (2011), 1–11.

[90] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. ACM, 459. https://doi.org/10.1145/3173574.3174033

[91] David A Puts, Benedict C Jones, and Lisa M DeBruine. 2012. Sexual selection on human faces and voices. *Journal of sex research* 49, 2-3 (2012), 227–243.

[92] Brigitte Roeder, Lisa Demuth, Judith Streb, and Frank Rösler. 2003. Semantic and syntactic priming in auditory word recognition in congenitally blind adults. *Language and Cognitive Processes - LANG COGNITIVE PROCESS* 18 (02 2003), 1–20. https://doi.org/10.1080/01690960143000407

[93] Brigitte Roeder, Wolfgang Teder-Sälejärvi, Annette Sterr, Frank Rösler, Steven Hillyard, and Helen Neville. 1999. Improved auditory spatial tuning in blind humans. *Nature* 400 (08 1999), 162–6. https://doi.org/10.1038/22106

[94] David A Ross, Ingrid R Olson, and John C Gore. 2003. Cortical plasticity in an early blind musician: an fMRl study. *Magnetic resonance imaging* 21 (2003), 821–828. https://doi.org/10.1016/s0730-725x(03)00103-6

[95] Tim Rowland. 1995. Hedges in mathematics talk: Linguistic pointers to uncertainty. *Educational Studies in Mathematics* 29, 4 (1995), 327–353.

[96] Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. 1986. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel* 81, 10.5555 (1986), 26853.

[97] Klaus R Scherer, Harvey London, and Jared J Wolf. 1973. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality* 7, 1 (1973), 31–44.

[98] Sajad Shirali-Shahreza and Gerald Penn. 2018. MOS Naturalness and the quest for human-like speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 346–352.

[99] M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. *Commun. ACM* 61, 3 (2018), 39–41.

[100] Vicki L Smith and Herbert H Clark. 1993. On the course of answering questions. *Journal of memory and language* 32, 1 (1993), 25–38.

[101] David Spiegelhalter. 2020. Should we trust algorithms? (2020).

[102] Amanda Stent, Ann K. Syrdal, and Taniya Mishra. 2011. On the intelligibility of fast synthesized speech for individuals with early-onset blindness. In *The 13th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '11, Dundee, Scotland, UK, October 24-26, 2011*. ACM, 211–218. https://doi.org/10.1145/2049536.2049574

[103] J. Trouvain. 2007. On the comprehension of extremely fast synthetic speech.

[104] Patrice Voss, Maryse Lassonde, Frédéric 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 : CB* 14 (11 2004), 1734–8. https://doi.org/10.1016/j.cub.2004.09.051

[105] Michael B Walker. 1977. The relative importance of verbal and nonverbal cues in the expression of confidence. *Australian Journal of Psychology* 29, 1 (1977), 45–57.

[106] Valerie SL Williams, Lyle V Jones, and John W Tukey. 1999. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of educational and behavioral statistics* 24, 1 (1999), 42–69.

[107] Eugene Winograd, Nancy H Kerr, and Melanie J Spence. 1984. Voice recognition: Effects of orienting task, and a test of blind versus sighted listeners. *The American journal of psychology* (1984), 57–70.