Blind and Low Vision Individuals' Detection of Audio Deepfakes

Filipo Sharevski
DePaul University
Chicago, IL, United States
fsharevs@depaul.edu

Aziz Zeidieh
Unviersity of Illinois
Urbana-Champaign, IL, United States
azeidi2@illinois.edu

Jennifer Vander Loop
DePaul University
Chicago, IL, United States
jvande27@depaul.edu

Peter Jachim
DePaul University
Chicago, IL, United States
pjachim@depaul.edu

Abstract

Audio deepfakes are a form of deception where convincing speech sentences are synthesized through machine learning means to give an impression of a human speaker. Audio deepfakes emerge as an attractive vector for targeting users that rely on audio accessibility, such as individuals who are blind or low vision. The critical reliance on speech both as a medium and an affordance puts this population at an undue risk of being deceived as they rely solely on themselves to detect whether a piece of audio is a deepfake or not. To better understand the nature of this risk considering the nuanced reliance on assistive technologies such as screen readers, we conducted a user study with n=16 blind and low vision individuals from the US. Our participants achieved an overall discernment accuracy of 59%, and clips identified as deep fakes were only actually deepfakes in 50.8% of the cases (precision). The participants that self-identified as "low vision" performed slightly better (accuracy of 61%, precision of 64%) compared to the ones that self-identified as "blind" (accuracy of 55%, precision of 56%). Our qualitative results show that the participants in the "blind" group mostly considered a combination of infliction, imperfections in the voice, and the intensity in the speech delivery as discernment factors. The participants in the "low vision" group mostly used the speaker's pitch, enunciation, emotion, and the fluency and articulation of the speaker as discernment cues. Overall, participants felt that audio deepfakes have the potential to deceive visually impaired individuals with political disinformation, impersonate their voice in authentication and smart homes, and specifically target them with voice phishing and enhanced scams.

CCS Concepts

• **Security and privacy → Social aspects of security and privacy**; • **Human-centered computing → Empirical studies in accessibility**; • **Social and professional topics → People with disabilities**.

CCS '24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3690305

Keywords

Audio deepfakes, blind, low vision, users, perception, detection

ACM Reference Format:
Filipo Sharevski, Aziz Zeidieh, Jennifer Vander Loop, and Peter Jachim. 2024. Blind and Low Vision Individuals' Detection of Audio Deepfakes. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3690305

1 Introduction

Impersonating audio by means of deep neural network machine learning – or *audio deepfakes* – recently became a trivial endeavor. Expectedly, immediate concerns abounded about the potential use of audio deepfakes for various forms of deception. Incidents such as faking the LastPass CEO's voice as a pretext in a phishing campaign [22] or faking President Joe Biden's voice in an early electoral disinformation campaign [37] give pause to rethink the disruptive capacity a convincing, yet synthesized speech could have on people's safety, security, and overall wellbeing. Conjectures already see audio deepfakes used for blackmailing, intimidation, ideological influencing, and inciting violence [6].

Audio deepfakes could take many forms of deception, usually depending on the context where they are employed. In the aforementioned examples, the targets – a LastPass employee and North New Hampshire voters – are presumably aware of the speaker's voice features, so the deception is driven towards making them verify the speech as legitimate relative to how plausible the context is, e.g., an urgent CEO request or voting in primaries. But in other examples, such as unsolicited calls or call center conversations, targets possess little knowledge about the speaker's voice features, so the deception is neither impeded by the speech verification task nor by inferring a plausibility of a context, given that it is easy to create a believable pretext. Machines are not spared from deception as using a cloned voice to trick a home voice assistant [1] or a biometric authentication system [24] is easy to perform [19].

People, nonetheless, remain the main targets as audio deepfakes are the ideal vector for social engineering campaigns. Work focused on social engineering deception shows that a persuasive pretext from an authoritative source revolving around scarcity tricks people into believing that email/SMS senders or voice callers with continuing effectiveness [36]. Audio deepfakes, like never before, allow social engineers to sample, scale, recycle, and combine large numbers of authoritative speakers that verbalize a persuasive scarcity

pretext with unprecedented ease. It comes as no surprise that the Federal Communication Commission (FCC) considers audio deepfakes an illegal form of voice communication [9]. Audio deepfakes also offer an easy way to seed disinformation that targets people for political [15] or information warfare purposes [4].

The threat of widespread deception of targeting people through audio deepfakes became apparent quickly, and work commenced to assess the ability of ordinary people to discern between legitimate and illusory speech. Challenges, such as the ASVspoof2021 [24], are continuously organized where human listeners are attempting to classify audio samples as deepfake or legitimate (*bona fide* speech). Academic studies too involve humans in similar tasks where individuals test their ability to discern deepfakes from legitimate ones [25, 27]. Researchers also produce datasets of audio deepfakes such as WaveFake [12] or FakeAVCeleb [21] to encourage in-depth exploration of the nature of the audio deepfake deception and the associated risk to people in everyday life.

However, these efforts rarely involve people who are blind or low vision. This is rather a curiosity, given that blind and low vision individuals – unlike their sighted counterparts – use speech not just as a medium but also as an accessibility affordance. Audio deepfakes, so far, pertain only to the medium part because the accessibility affordance might have been considered a form of an audio deepfake (e.g., text-to-speech synthesis). But the accessibility affordance – be that a voice-over or a screen reader – might also be manipulated to convey an incorrect meaning of the text [32]. This double exposure makes a compelling case to understand how blind and low vision people assess potential audio deepfake deception. Given that this population's visual perception falls on a spectrum, the reliance on speech as a medium also varies as a factor in spotting deception (e.g., a blind person might entirely rely on a screen reader and a low vision person might use a screen magnifier in addition). The varying degree to which speech as a medium is utilized from an accessibility perspective also makes a further case to explore whether one's particular nature of the visual impairment factors in the way they discern audio deepfakes from legitimate voice. To address these gaps, we conducted semi-structured interviews with 16 blind or low vision individuals, answering:

- **RQ1:** How accurate are blind and low vision individuals in discerning legitimate from deepfake audio?
- **RQ2:** What cues blind and low vision individuals use when discerning legitimate from deepfake audio?
- **RQ3:** Is there a difference in the accuracy and the cues used to discern legitimate from deepfake audio between blind individuals and low vision individuals?
- **RQ4:** What perspectives do blind and low vision individuals have relative to the deceptive ability of audio deepfake pertaining to this vulnerable population?

We obtained approval from our Institutional Review Board (IRB) to conduct an open-ended interview user study with a sample of n=16 individuals who are low vision (n=8) or blind (n=8) (original approval for 12 participants or above, 6 of each visual disorder self-identification) who are aware of the ability to synthesize a convincing speech through machine learning means. We set to perform our observations over Zoom, where we shared an audio-only part of a screen from a browser that had a survey set up with embedded audio clips, offering a remote function to play it and select a choice whether the audio is a (i) deepfake or (ii) legitimate. There were 16 audio clips in total that participants were exposed to, and each participant assessed them in a randomized order (additionally, the choices of deepfake/legitimate were randomized, too).

Eight of the clips – four legitimate and four deepfake – were carefully selected from the ASVspoof2021 dataset [24] (the curation process in elaborated in Section 3). The other eight were created by asking an arbitrary selection of four speakers to read several sentences selected from the VCTK Corpus (this was used to seed the ASVspoof2021 data set, too) [39]. We then used these spoken sentences as input in ElevenLabs to create four deepfake sentences, also using the VCTK corpus [7]. During each detection task, we asked participants to verbalize the cues in the clip as well as any prior experience that helped them determine their decision about the clip's provenance. Once the participants completed the detection tasks, we asked them about their general perspectives on the deceptive ability of audio deepfake pertaining specifically to blind or low vision individuals. We then shared their score in the detection task and asked for their perspectives on their performance in order to contextualize the deceptive nature of the audio deepfakes.

We found that the blind and low vision participants in our study were accurate in 59% of the cases in discerning deepfake from legitimate audios. Our participants performed slightly better with the spoofed speech from the ASVspoof2021 data set (63%) than with the speech we generated using the ElevenLabs algorithm (54%). They correctly identified 60.2% of deepfakes (recall), and audio clips identified as deepfakes were correctly identified 50.8% of the time (precision), for an overall detection reliability F_1 score of 55.1%. Controlling for the visual disorder or the full/partial reliance on screen readers, we found a slight advantage in discernment of the participants that self-identified as "low vision" (accuracy of 62%) compared to the ones that self-identified as "blind" (accuracy of 55%). Participants in the "blind" group correctly identified 50% of deepfakes (recall), and audio clips identified as deepfakes were correctly identified 56% of the time (precision), for an overall detection reliability F_1 score of 53%. Participants in the "low vision" group also correctly identified 50% of deepfakes (recall), and audio clips identified as deepfakes were correctly identified 64% of the time (precision), for an overall detection reliability F_1 score of 56%.

The participants in the "blind" group showed better assessment performance than the "low vision" group when assessing four clips out of all 16 we used as stimuli: one legitimate and three fake. The "low vision" group performed better when assessing eight clips (five legitimate and three fake), and both groups were tied in the performance for four clips (two legitimate and two fake). Cue-wise, our participants were less concerned with the recording quality and more so relied on the presence/absence of cues such as (i) inflection, speech imperfections (notably, breathing noises), and the subjective impression of the speech intensity and speaker's identity ("blind" group); or (ii) emotion, enunciation, and fluency in expression and articulation ("low vision" group). Concerned by the audio deepfakes' deceptiveness, our participants felt that the convincing quality of voices they rely on – both as content as well as an accessibility support – warrants rethinking as blind and low vision individuals could unduly be targeted with political disinformation, enhanced scams and voice phishing, and voice authentication impersonation.

2 Background and Related Work

2.1 Audio Deepfakes and Humans

Deepfake audio emerged as a more subtle way of deception compared to deepfake videos [2], given that humans can only rely on familiarity with the speaker's voice or the context in which the speech was delivered. President Biden was impersonated in an audio deepfake in a political disinformation campaign, and the voice of the security firm LastPass CEO was used as a pretext in a voice phishing campaign [22, 37]. These are two examples that highlight this subtleness; in both cases, convincing speech understandably leads to conjectures of possible deceptiveness, especially when the speaker's voice is unknown, or the speaker is not a high-profile person. Despite the concerning nature of this "exposure," few studies, shown in Table 1, so far attempted to understand the ability of humans to discern a legitimate audio from a deepfake one.

Table 1: Audio Deepfakes and Humans

Author et al.	Participants	Stimuli (l/f)	Dataset(s)	Accuracy
Müeller [27]	472 (GER)	5/5	ASVSpoof2019	80%
Mai [25]	281 (USA), 248 (CHN)	10/10	LJSpeech, CSMSC	73%
Frank [11]	822 (USA)	15/15	LJSpeech	50.57%
	875 (GER)	15/15	HUI	59.15%
	912 (CHN)	15/15	CSMSC	51.73%
Han [16]	12 (USA, Blind)	30/33	ASVspoof2021	57%
	30 (USA, Blind)	40/56	ASVspoof2021	59%
	30 (USA, Sighted)	40/56	ASVspoof2021	59%

Comparing humans' and machines' abilities to detect deepfake audio, Müeller et al. used the ASVspoof2019 dataset (7355 legitimate stimuli and 4919 deepfake) [29] and found that humans and automated methods face the same weaknesses when detecting audio deepfakes in real-world scenarios (simulated through a game where each participant evaluated a random selection of five legitimate and five deepfake stimuli). People's detection performance of audio deepfakes was found to peak at 80% accuracy, even when receiving feedback on their performance and being allowed to use it to improve their scores [27]. Mai et al. conducted a study using 20 audio stimuli (10 legitimate and 10 deepfake) in both English and Mandarin, equally balanced and generated using the LJSpeech and CSMSC datasets, respectively. Their participants were only able to correctly identify the deepfake audio 73% of the time (aggregate), mostly relying on a subjective impression of the "naturalness" of the speech [25]. Frank et al. [11] compared the performance of English (US), German, and Mandarin individuals to detect a set of 30 audio stimuli (15 legitimate and 15 deepfake, generated using three TTS datasets: LJSpeech – English, CSMSC – Mandarin, and HUI – German) in their native language and found that the participants from the US demonstrated an average accuracy of 50.57%, the ones from Germany 59.15%, and the ones from China 51.73%.

The only study to date that compared the audio deepfake detection performance between blind and sighted individuals, using the ASVspoof2021 dataset, was conducted by Han et al. [16]. They found that both groups of participants performed about the same, with an overall detection accuracy of 59% for both populations. The blind individuals were better at detecting text-to-speech (TTS) generated deepfakes audio, while the sighted individuals were better at detecting Voice Cloning (VC) deepfake audio. All of their participants revealed that they use human traits in audio, such as accents, vocal inflections, breathing patterns, and emotions, as cues of deception. When explaining their method of discernment, the blind participants tended to use their experience with screen readers as a basis of comparison. In contrast, the sighted participants used their experience with authentic human voices.

2.2 Audio Deepfakes and Humans Who Are Blind or Low Vision

So far, the academic attention has been focused on determining the accuracy with which humans discriminate between a legitimate audio and a deepfake one. This performance, as shown in Table 1, is either benchmarked between humans and machines [27], languages [11, 25], or presence/absence of vision [16]. In the studies with only sighted individuals, the accuracy was tested only as a quantitative measure without an effort to provide a qualitative analysis of the decision-making process behind the discernment. This is a limitation that leaves the realm of cues of deception unexplored, offering no actionable outcomes or possible solutions that could help humans better respond to audio deepfakes. Even quantitatively, the studies are also restricted to accuracy as a performance measure, without disclosing details about the precision (the number of clips identified as fake that are in fact fake), recall (correct number of fake clips identified), reliability (i.e., the F_1 score or avoiding incorrect decisions), and specificity (correctly identifying legitimate clips).

Another limitation that is noticeable in the studies with sighted individuals is the use of audio stimuli from datasets that are generated no later than 2019. While the use of public datasets demonstrates a methodological rigor and allows for later comparison, it nonetheless limits the test of advanced deepfake techniques that are developed past 2019 (mostly the audio deepfakes in these datasets are TTS generated with a limited voice cloning samples). The only study that compares the accuracy between humans based on their visual disorder (presence/absence of vision) and not on language [16] attempts to rectify this limitation in that it uses the updated ASVspoof2021 dataset. But even with this dataset, all the studies fail to measure the ability of humans to discern a legitimate audio from a fake one when its spoken by random, unfamiliar speakers and it is generated by an available synthetic-voice cloning technology that was already used in the real world to deceive humans [28].

Han et al. [16] are also the only study that adds a qualitative dimension to the quantitative assessment of the accuracy in deepfake detection between blind and sighted individuals. While this is a much-needed step towards assessing the susceptibility to audio deepfakes of a particularly vulnerable population such as blind individuals, it leaves out the population of low vision individuals that are, similar to their blind counterparts, reliant on assistive technology such as screen readers [23, 34]. As visual disorders could be on a spectrum, the nuanced understanding of how individuals that are solely reliant on screen readers perform compared to ones that partially use them is equally important for more detailed and inclusive mapping out of the realm of cues of deception. In this

context, a set of further limitations of the Han et al. [16] study are the imbalanced set of legitimate/fake stimuli, the biased selection of only the top false positive and top false negative audio clips from the ASVspoof2021 (as an output from a machine classifier), as well as the large number of stimuli used in the assessment.

The imbalanced set might allow for accuracy tests, but it might restrict a more detailed understanding of the reliability and specificity in the detection (neither of which were reported in the study). The selection of the most deceptive clips helps understand the situations where an individual – blind or sighted – might be tricked with the highest probability, but it fails to account for realistic situations where the cues for deception are in fact helpful in correct discernment. This knowledge is helpful when considered in creating tailored detection or awareness-raising solutions for blind, low vision, or sighted individuals (also absent from this study). The large testing set of 63 audio clips might not be representative of the true discerning ability due to survey/interview fatigue, especially for blind individuals that need additional time to access the recording and submit their response even if the clips lasted not more than 10 seconds each (on par with the other assessment studies). Table 1 reveals that the studies with sighted individuals balanced and limited the number of stimuli to not more than 30 clips in total to safeguard for this potential threat to results reliability. Aside from these methodological limitations the Han et al. [16] study only provides insights into the detection decision-making process without probing the participants to contextualize the study experience and offer their perspectives on how audio deepfakes could affect or are already affecting their real lives (a line of inquiry that is also absent from all the other studies with sighted individuals).

3 Study Methodology

To address the aforementioned limitations, we created a study to test the ability of blind and low vision individuals to tell a fake audio from legitimate audio (that is, test their discernment performance relative to accuracy, precision, recall, reliability, and specificity); understand their discernment approach; and learn about their concerns relative to this new type of deception. The goal of the study was not to necessarily generalize the results to the entire population of blind or low vision users nor to derive any strict commonalities in the discernment approach, per se. Instead, our motivation was to explore the engagement with this type of medium from a susceptibility perspective and compare the exposure between individuals that – per their self-identified visual disorder status – solely or partially rely on synthetic voice as an avoidance (screen reader).

3.1 Ethical Considerations, Risks, Safeguards

As this was a non-full disclosure study that employed deception, even if minor and revealed, it was important for us to establish trust and assurances about the goals of the study and the safeguard protections we had in place. Following the suggestions for considering the ethical aspects when doing research with blind or low vision individuals from prior work [17], we first obtained verbal consent both before we started the Zoom audio recording of the interview and afterward (to have evidence in our transcripts, but also to avoid creating a recording in case a potential participant does not consent, in which we would have thanked them and closed the Zoom session). Next, we communicated that the goal of our study was to capture the "richness" of their experiences and decision-making process around deepfakes rather than simply benchmarking their detection of an arbitrary selection of audio clips. We emphasized that we deliberately selected audio clips that have no direct relation to their everyday encounter of speech and that the accuracy assessment we ask them is done as a basis for building resilience against audio deepfakes among this population, both in deliberating their deceptive ability and developing usable detection interventions and awareness training for blind and low vision individuals.

We did this to avoid leaving our participants worse than they were before participating and feeling that they were unduly deceived and concerned about audio deepfakes in the future. In addition, we notified our participants that we designed the study to be inclusive in that it considered their basic capabilities (e.g., minimum requirements for internet connection and interview technologies, technology assistance, own choice of accessibility technology, own selection of time of participation and environment) and that it was relatable to them and conducted in a way that was non-judgmental and empathetic. We also offered the option for them to choose not to participate and abandon any question at any point in time if they felt like doing so. We also allowed them to replay each audio clip as part of the assessment and deliberation relative to the cues they used in discerning its authenticity. To avoid contamination and noise in the results, we didn't allow participants to go back and change their answers, though we noted that they are free to use as much time as they need to meaningfully engage with each clip before verbalizing their discernment process and venturing a decision. Prior to doing any of the assessment tasks and the interview, we told them they could ask us to stop the interview, stop the recording, or remove any answers or readings at any point in time.

Only after we received the participants' explicit agreement for the assessment structured interview did we commence the audio-only recorded Zoom session and proceed to complete the tasks and the interviews. Following the recommendations for doing usable security research with blind and low vision users [33], we allowed them to verbalize the process; to give comments, complaints, and suggestions; and to verbalize any other experiences that were not necessarily with the selected clips but with synthesized/real speech or audio deepfakes – this was done in order to allow for them to fully express the way their everyday interaction with speech or multimedia content shapes the way they gauge the provenance of an arbitrary audio. After we collected participants' assessments and recorded their deliberations as part of the interview - both relative to the clips and their perspectives about the audio deepfakes' general deceptive ability, we verbosely debriefed them about the study. First, we disclosed that half of the clips they heard were audio deepfakes – four were taken from the ASVspoof2021 [24] dataset and four of them were created by us, the researchers, using the most popular audio deepfake generator, ElevenLabs [7]. We employed lengthy descriptions of the nature of the ASVspoof2021 detection challenge, its goal to improve the overall detection of audio deepfakes, and the ability of humans to correctly discern natural speech from a convincingly synthesized one. We also explained our decision to use ElevenLabs as the leading synthetic-voice cloning technology used to create the convincing, yet deepfake speech of President Biden for disinformation purposes [28].

We assured our participants that we were not involved with ASVspoof2021, ElevenLabs, or any technology that is used to produce audio deepfakes they heard during the study or might have heard in the past. We were also careful not to appear in favor nor support of the creation or dissemination of audio deepfakes for any purposes – given the freely available tools online to do so – both to maintain full researcher impartiality and to avoid inspiring unintended deceptive practices as a result of the study participation. We communicated that our ultimate goal is to create meaningful *inclusion* for blind and low vision individuals in developing detection and awareness solutions, as this vulnerable population is often excluded when such interventions are developed for other types of deception (e.g., phishing emails). We also indicated that there is a possibility that after the publication of the results of the study, one could misuse or misinterpret them in the broader online deception context. That is, an adversary could use audio deepfakes, in particular the assessment results and their discernment approach, to target blind and low vision individuals,

We had no available safeguard against this risk, and we offered participants the option to withdraw their participation if they wished. Here, we stated that we strongly condemn such behavior and recommended that they contact us if they experience or suspect any such activity at any point in the future for any help or assistance if they deem necessary. We nonetheless pointed out participants to general voice phishing resources where audio deepfakes have already been used for pretexting targets if they wished to further raise or check their awareness [5]. We also pointed out the general resources about voice cloning communication and protections regularly updated by the FCC [8]. We obtained a correct understanding of their experiences, discernment decision-making, and perspectives. We reviewed the main points recorded during the interview and clarified any misunderstandings we might have. We also sent a draft of our paper to our participants for feedback.

3.2 Participant Recruitment

For our recruitment, we followed Gerber's advice when doing usability and accessibility studies with blind and low vision people [13]. Accordingly, we recruited participants who use a screen reader, screen magnifier, or both and who regularly use web/email clients and could obtain online information aurally without problems. They had to be individuals 18 years of age or older, from the US, who had internet access on their own device, client, and browser, and were English-speaking and literate. As one of the researchers is a legally blind individual, we recruited potential participants through snowballing, where we partially sampled personal acquaintances and partially a pool of blind and low vision participants recommended by one of our acquaintances. We used a formal recruitment email approved by our IRB (Appendix [35]). We arranged audio-only recorded Zoom interviews with interested respondents on a rolling basis, requesting that each participant has access to their preferred way of using emails and their assistive technology.

The potential participants were informed that the interview is audio-only, on Zoom, recorded for transcription purposes only (audio immediately deleted), and that there is no requirement for them to use their real name during the participation (choose an alias to preserve their anonymity). The interviews lasted on average 60 minutes, and we compensated each participant with either a $25 paper Amazon Gift card (card number and code were read out loud at the end of the interview after the recording was stopped so the participants could redeem it at the conclusion of participation on their own) or an Amazon eGift card (voluntarily provided email for delivery that was not recorded, collected, or included in any way in the analysis of the results to preserve anonymity). Each interview was done with open-ended questions (Appendix [35]).

We concluded our recruitment with a sample of 16 participants as we reached a thematic saturation, i.e., we collected data up to the point where there were fewer surprises in the responses to the RQs and no more emergent thematic patterns. As we worked with at-risk participants from a smaller population than the population of sighted individuals, we employed the similar approach taken [16] where the assessments were performed with a comparable number of individuals with visual disorders (12 and 30, respectively). We also followed the approach in [23] where investigating visually impaired users' security experiences usually involved 10-15 participants. The demographics, including the participant's visual disorder self-identification, are given in Table 2.

Table 2: Participant Demographics

P#	Gender	Age	Education	Self-Identification
P1	Female	30	Post-Graduate	Blind
P2	Female	30	Post-Graduate	Blind
P3	Female	30	College	Blind
P4	Male	24	College	Blind
P5	Female	42	Post-Graduate	Low Vision
P6	Female	26	Post-Graduate	Low Vision
P7	Male	64	Post-Graduate	Blind
P8	Female	38	College	Low Vision
P9	Female	20	College	Low Vision
P10	Male	30	College	Low Vision
P11	Female	18	High-school	Low Vision
P12	Female	27	College	Blind
P13	Male	32	College	Low Vision
P14	Male	41	College	Blind
P15	Male	40	College	Blind
P16	Male	33	College	Low Vision

3.3 Study Stimuli

Past studies focused on human evaluation of audio deepfakes exclusively use *known datasets* of synthesized speech [16, 27], developed for this particular purpose. For half our study stimuli, we also followed this practice and used the ASVspoof2021 dataset, the Logical Access (LA) attacks segment (spoofing remote authentication, e.g., telephone banking service), and the Deep Fake (DF) attack segment (manual speaker verification). We didn't use the Physical Access (PA) attacks segment of the dataset (spoofing in-person authentication through speech, e.g., door-mounted voice/speech authentication) because we aimed to assess how human listeners (in our case, blind or low vision individuals), instead of machines, discern an audio deepfake from a legitimate one. We first singled out only the US English speakers from the dataset, selecting two female and two male speakers, with a total of 5576 audio clips (*bona fide* or "legitimate" and spoofed or "deepfake" speech included).

Though each audio clip lasted around four seconds, the total number of audio clips was prohibitive to be tested with our sample due to limitations in time and compensation for participants as well as avoiding negative effects such as habituation.

We opted for a curation as a viable alternative that fitted the way blind and low vision people habitually participate in evaluation studies with content that is relevant to them and their unique everyday experience with synthesized speech as a result of using assistive technologies [33]. We first decided on a 95% confidence interval and a desired margin of error of 5% of the primary selection, yielding a preliminary known set of stimuli of 360 audio clips. This number was still prohibitive for testing, and we reviewed each recording to include only audio clips that contained understandable utterances as-is and excluded recordings that might be confusing or missing context for US listeners. As the ASVSpoof2021 is based on the VCTK Corpus, the audio clips contain 400 sentences selected from a newspaper (The Herald Glasgow), the rainbow passage, and an elicitation paragraph. Many of these clips contain references to past events or cultural aspects of local relevance (e.g., an interview with a soccer coach from a match played a long time ago in the Scottish First Division) that are of little to no pertinence to our study population, so we excluded such examples.

As we wanted to also analyze the decision-making process around the assessment and the general take on deepfakes, we determined that obtaining meaningful yet valid results would entail no more than 16 recordings in total (as the interviews were capped at one hour, we estimated (i) 3 minutes on average for a blind or low vision individual to play the audio clip, take time to process it, verbalize the discernment process and make a selection; (ii) 10 minutes to communicate back the assessment performance and discuss their take as well answer the RQ3; and (iii) 2 minutes variance for setting up their assistive technologies they). For the deepfake audio clips, we selected two different Text-to-Speech (TTS) spoofing algorithms (A07, A09), one combined TTS and Voice Conversation (VC) spoofing algorithm (A15) and one (VC) spoofing algorithm (A19) to control an explore how the exposure to TTS through assistive technologies factors in blind or low vision participants' detection approach. We could have opted to use several samples to cover the preliminarily selected clips (e.g., make ten partitions and test each one with three samples of at least ten participants), but it would have been impossible to recruit a large enough sample as well as infeasible to manage the participation overall.

We specifically wanted to test a *custom dataset* where we varied the deepfake technology but retained the basic VCTK corpus of speech, we further narrowed the number of ASVspoof2021 clips to eight (four are legitimate and four are deepfake, balanced for female and male speaker), given in Appendix [35]. For our *custom dataset*, we used the VCTK corpus, only we arbitrary selected two female and two male English speakers (only US accent) to create the bona fide or "legitimate" speech. This bona fide speech was used as an input into a current-of-the-shelf synthetic-voice cloning technology, ElevenLabs [7], to produce the deepfake speech for each speaker based on another two female and two male spoken sentences from the ASVspoof2021. We decided to emulate a scenario where any attacker might use readily available tools to generate a speech in order to trick a human listener – in addition to spoofing authentication – including voice phishing or spam [7]. This was inspired by the use of the very same technology for disinformation purposes (spreading misinformation with an intent to deceive) in cloning President Biden's voice [28]. The resulting half of the custom-created study stimuli is given in Appendix [35].

3.4 Data Collection

Before collecting the data, we established a baseline for each of the tasks and questions that our participants were invited to answer. For the first research question, we defined accuracy as the "the closeness of the participant's discernment decision of the nature of the audio clip – deepfake or legitimate – to the actual nature of the audio clip (known to the researchers, but unknown to the participants)." Many human detection tasks, including ASVspoof2021, of evaluate the "naturalness" of the speech as a Mean Opinion Score (MOS) on a scale ranging from 1–Very Unnatural to 5–Very Natural. Doing such baselining was deemed complex given that the participants in our sample habitually encounter varying degrees of both natural and unnatural speech through their use of assistive technologies, so we opted to use a simple categorical variable of 0 –deepfake, 1–legitimate to avoid misinterpretation of the main task at hand.

For the second research question, we pointed out to participants that we refer to "cues" in the audio media as "any indicator in the recording, the speech delivery, or the combination of both that – based on their subjective experience – reveals a particular quality, be that a synthesized or a legitimate speech." To make further clarifications, we noted that these indicators could be related to the speaker's expression, speech delivery, the quality of the recording, or the recording identity [14, 31]. For the third research question, we defined the deceptiveness of audio deepfakes as "the ability of an audio clip to create a perception of a convincing, natural speech when, in fact, the speech is created through automated means." To avoid any social desirability bias, here, we emphasized that the interview questions are permissive of the sensitive attitudes, beliefs, or behaviors that shaped the participants' perspectives and noted that we don't have a predefined expectation of their answer.

Once we had the baseline in place, we shared a screen and the audio embedded in a Qualtrics survey over Zoom and offered the participants to either use their remote control to interact with the survey or provide answers and let us act on their behalf. Each audio clip was implemented as a single question on a single page, and the questions were randomized per participant to avoid ordering bias. Participants had the opportunity to replay the audio clips if they wanted and were encouraged to use any audio equipment they usually use to create realistic conditions where blind or low vision individuals interact with audio and assistive technologies in their everyday lives. On average, we planned for roughly 3 minutes of overall assessment time per audio clip. After that, we calculated their assessment performance, shared it with our participants, and asked their perspectives on it and the general deceptiveness of audio deepfakes in the remaining hour-long participation.

The assessment answers were recorded directly in Qualtrics, and the remaining data was included in the interview transcripts. Initially, the interview transcripts from our Zoom sessions were not anonymized, but we removed any names and references to individual participants and deleted the audio recordings altogether. The transcripts, assigned only with a participant number in the order

of participation, were stored on a secure server that only the researchers had access to. Each interview was done with open-ended questions listed in the interview script (Appendix [35]). We concluded our data collection with a sample of 16 blind or low vision participants as we reached thematic saturation (i.e., we collected data up to the point where there were fewer surprises in the responses to the research questions and no more emergent patterns). As part of the debriefing process (Appendix [35]), participants were offered the option to withdraw from the study after finding out about the nature of the recordings and their performance (none of them did, and in fact, all participants felt the interview was beneficial to them in raising their audio deepfake awareness).

3.5 Data Analysis

We used basic descriptive statistics to calculate the performance of the deepfake/legitimate audio discernment tasks. With the interview data, we first performed a round of open coding for two, arbitrarily selected participants to familiarize them with the responses and derive codes directly from them. Then we internally discussed the individual coding schemes and together converged on an agreed codebook (Appendix [35]). Two researchers used this codebook to independently code the remaining interviews. The inter-coder agreement (or inter-rater reliability) or average Cohen's Kappa coefficient (κ) for all themes in our data was 0.869 (a value above 0.75 is considered as an excellent agreement [10]). The themes we identified were then discussed and interpreted, and an example quotation was selected to represent each of the findings. We engaged with the results following the reflexive thematic analysis approach to examine the relationships between themes and draw together a narrative, as well as re-interpret them in the socio-technical context of audio as a dual-use technology for people with visual impairments within which they emerged.

For the discernment cues (linguistic and acoustic expression, recording quality, and recording identity) we assessed the thematic saturation relative to the general quality assessment aspects of digitally recorded audio [14, 31]. For the perspectives (personal performance, and personal safety and security) part of the RQ4, we acknowledge that the claim to saturation might be problematic as the perspectives of blind or low vision individuals on the audio deepfakes' deceptive nature is practically limitless [30]. We also acknowledge this issue in the limitations section, but we did however assessed the thematic saturation in the context of the audio deepfake attacks currently spotted in the wild relative to disinformation [22, 37], as well as phone scams, voice authentication, and voice impersonation [36]. The thematic saturation mainly pertained to the RQ2 and the qualitative part of RQ3. Relative to RQ1 and the quantitative part of RQ3 we acknowledge that the assessment of fake audio detection does not rely on thematic saturation.

4 Results

4.1 RQ1: Detection Performance

The accuracy performance achieved by our participants is given in Table 3. Overall, our sample achieved a 58.6% accuracy over the entire set of study stimuli, suggesting that blind or low vision individuals do perform better than chance, though they still are concernedly susceptible to an audio deepfake deception. Broken down per segment of stimuli, our participants performed better with the ASVspoof2021 (63%) stimuli than with the ElevenLabs (53%). The highest overall accuracy of 69% (5 total misidentifications) per participant was achieved by four of them (**P10, P12, P13, and P15**), while the lowest accuracy per participant of 44% (9 misidentifications) was achieved by two participants (**P1** and **P7**).

Relative to the accuracy per legitimate audio clip, our participants performed much better with the ElevenLabs part of the stimuli, achieving in both cases (clips **E1** and **E3**) a 94% accuracy and once an 88% accuracy (clip **E3**). Here, we noticed a sizable drop in accuracy was for clip **E2** to 38% (as shown in Appendix [35]), the participants' perception was that the speaker lacks emotion and intensity, the recording was flawless, and the delivery was more close to machine-generated than a natural voice). For the legitimate part of the ASVspoof stimuli, our participant's best accuracy (75%) was achieved for clip **A3** and the worst accuracy (31%) for clip **A4** (also shown in Appendix [35], the perception was that the speaker articulation was close to reading, with a monotone voice, a lack of background noise, and lack of natural delivery).

Relative to the accuracy per deepfake audio clip, our participants performed much better with the ASVspoof2021 part of the stimuli, achieving a 100% accuracy for clip **A6** (generated using the A09 algorithm – a recurrent neural network TTS using the Vocane vocoder [40]) and a 69% accuracy for clips **A5** (generated using the A07 algorithm – another recurrent neural network TTS using the WORLD vocoder, but post-processed by WaveCycleGAN2 for better natural sounding [3]) and **A7** (generated using the A15 algorithm – a combined TTS and VC using the WaveNet vocoder [20]). The lowest accuracy of 50% was achieved for the **A8** clip, generated using a transfer-function-based VC system [26] (as shown in Appendix [35]), half of the participants felt the speaker's articulation was distinctly natural, given the presence of an inflection and a varying pitch characteristic for authentic speakers).

The detection accuracy performance of our sample noticeably dropped for the audio clips generated using the ElevenLabs algorithm. Here, the best performance was for clip **E7** (50%) where half of the participants (Appendix [35]), felt that the speaker's fluency, inflection, intensity, and expressed emotion were sufficient cues that the clip is real. The next best performance was for clip **E6**, where a slight majority of the participants thought it was real because it had a natural, human-sounding pitch and noticeable inflection. The second to worst accuracy performance was for clip **E5** where all but 19% of the participants felt fluency of the speaker and the natural sounding pitch were sufficient cues it was a real person speaking. The lowest performance of only 6% was for clip **E8** where all but one participant were under the impression that it was a real articulated speaker because the voice had a naturally sounding pitch, as shown in Appendix [35].

The blind and low vision individuals in our sample, as shown in Table 4, had an overall precision of 60.2%, meaning that 60.2% of the clips identified as fake were in fact fake, demonstrating a much better precision with the ASVspoof2021 clips (71.9%) compared to the ones created with the ElevenLabs algorithm (57.6%). Recall-wise, our participants performed comparably in both sets of stimuli, correctly identifying 60.2% of audio deepfakes in 60.2%. Judging by the differences in the F_1 score, our participants were far less reliable when it comes to assessing the ElevenLabs audio deepfakes (39.2%)

Table 3: RQ1: Accuracy Performance: Legitimate vs. Fake

	ASVspoof2021 – Overall Accuracy 63%								ElevenLabs – Overall Accuracy 54%								
P	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**Acc**
1	Legit	Fake	Fake	Fake	Legit	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Legit	Legit	44%
2	Legit	Legit	Legit	Fake	Legit	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Legit	Legit	Legit	Legit	56%
3	Fake	Legit	Legit	Fake	Fake	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Legit	Legit	56%
4	Fake	Legit	Legit	Legit	Fake	Fake	Fake	Legit	Fake	Legit	Fake	Legit	Legit	Legit	Fake	Legit	56%
5	Fake	Fake	Legit	Legit	Legit	Fake	Legit	Legit	Legit	Fake	Legit	Legit	Fake	Fake	Legit	Legit	50%
6	Fake	Legit	Legit	Fake	Fake	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Fake	Legit	63%
7	Fake	Fake	Legit	Fake	Legit	Fake	Fake	Legit	Legit	Fake	Legit	Fake	Legit	Fake	Legit	Fake	44%
8	Legit	Fake	Legit	Fake	Fake	Fake	Legit	Legit	Legit	Fake	Legit	Legit	Fake	Fake	Fake	Legit	63%
9	Legit	Legit	Fake	Fake	Fake	Fake	Fake	Legit	Legit	Legit	Legit	Legit	Legit	Legit	Fake	Legit	63%
10	Legit	Legit	Legit	Legit	Fake	Fake	Legit	Legit	Legit	Legit	Legit	Legit	Legit	Fake	Legit	Legit	69%
11	Legit	Fake	Legit	Legit	Fake	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Legit	Legit	63%
12	Legit	Fake	Legit	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Legit	Fake	Fake	Legit	69%
13	Legit	Fake	Fake	Legit	Fake	Fake	Fake	Legit	Legit	Legit	Legit	Legit	Legit	Fake	Legit	Legit	69%
14	Legit	Fake	Legit	Fake	Fake	Fake	Fake	Legit	Legit	Fake	Fake	Legit	Legit	Legit	Fake	Legit	50%
15	Fake	Legit	Legit	Fake	Fake	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Fake	Fake	Legit	69%
16	Legit	Legit	Fake	Fake	Legit	Fake	Fake	Legit	Legit	Fake	Legit	Legit	Fake	Legit	Fake	Legit	56%
Acc	63%	50%	75%	31%	69%	100%	69%	50%	94%	38%	88%	94%	19%	44%	50%	6%	
	Color Scheme: Light Violet – Legitimate, Light Yellow - Fake, Gradient Gray - Accuracy **Color Gradient**: Proportional to the accuracy percentage																

compared to the assessment of the ASVspoof2021 deepfakes (66.2%). Our participants were also slightly better in correctly selecting a legitimate audio when assessing the ASVspoof2021 clips (Specificity = 66%) compared to the ElevenLabs clips (Specificity = 52.6%).

Table 4: RQ1: Remaining Performance

Metric	ASVSpoof2021	ElevenLabs	Overall
Accuracy	0.633	0.539	0.586
Precision	0.719	0.297	0.508
Recall	0.613	0.576	0.602
F_1 **Score**	0.662	0.392	0.551
Specificity	0.660	0.526	0.574

4.2 RQ2: Detection Approach

The thematic analysis of the decision-making approach shared by our participants when discerning deepfake from legitimate audio clips revealed that participants rely on several groups of cues pertaining to the linguistic expression of the speaker, the acoustics, the recording quality, and the subjective impression of the recording identity. As shown in Appendix [35], our participants were less concerned with the recording quality and more so relied on the presence/absence of inflection, emotion, fluency in expression and articulation, as well as changes in pitch, intensity, and speech imperfections (notably, breathing noises) coupled with their subjective impression of the recordings' identity. We found a considerable difference, however, in how these cues were employed in the assessment of each stimuli group.

4.2.1 ASVspoof2021. When accessing the **A1** clip, the participants who correctly identified it as legitimate relied on imperfections such as "*the breathing at the end and the inflection clearly placed an end of the sentence*" (**P11**). Those that accessed the **A1** incorrectly as a fake felt that the clip "*sounded like it was taken from a training module, without much intensity and normal fluency*" (**P6**). The correct assessments of the **A2** clip mostly used the "*inflection and pacing (pitch)* **P10** of the speaker, while the incorrect ones were under the impression that "*it just sounded very emotionless*" (**P12**). The giveaway that the **A3** clip was correctly legitimate was the distinct "*background noise and static*" (**P3**) as well as imperfections such as "*swallowing*" (**P14**). What made our participants say otherwise, was mostly related to the feeling that the speaker was "*narrating without showing much emotions*" (**P9**). The correct assessment for the **A4** clip was predicated on background noise, though here a decisive cue was the speaker's "*intonation and the rise and fall of their voice*" (**P10**)

Assessing the **A5** clip, many of the participants got the impression that "*there is a robotic undertone in it*" (**P15**) to correctly discern it as a deepfake. Those who felt the voice was legitimate were predominately cued by the "*up and down voice pitch*" (**P7**) of the speaker. The **A6** clip was the only one that all the participants correctly said it was wrong, clearly tipped off by the "*obvious distortions*" (**P12**) and "*exact same tone and no inflection*" (**P7**) that made it "*sound like a screen reader*" (**P3**). The **A7** clip was also correctly deemed fake as participants felt it sounded "*robot-like, synthesized*" (**P1**), though those that felt it was legitimate felt the speaker "*articulates the sentence with a distinct inflection*" (**P2**). The split between the assessment of the **A8** clip was along the lines of "*flat, computer sounding pitch*" (**P11**) in the correct case and the "*absence of inflection*" (**P8**) in the incorrect one.

4.2.2 ElevenLabs. The legitimate **E1** clip was correctly assessed by all but one participant mostly because it "*had a distinct inflection*" (**P2**), a "*varying human pitch*" (**P8**) and one could easily hear the "*speaker breathing* (**P11**). Only participant **P4** was "*skeptical because there was a lack of fluency.*" For the **E2** clip, the majority of the

participants incorrectly felt it was fake because it "*sounded robotic*" (**P6**) or had a "*flat pitch*" (**P8**). All but two participants correctly assessed the **E3** clip, feeling that "*the voice conveys an emotion*" (**P5**) and *sounds natural, like a voice of real people with background noise*" (**P8**). Those who were incorrect felt that "*there is not much enthusiasm about the sentence*" (**P4**). The **E4** clip was nearly unanimously deemed as legitimate because "*it had human-like presence*" (**P9**) and the speaker utters "*each word in a way that implies he understands them better than a robot just repeating them*" (**P11**).

The real challenge came with the ElevenLabs deepfake audio clips. All but three participants were wrongly convinced that the **E5** clip was legitimate, given that "*she fluently spoke the words like she meant them*" (**P11**) and there was a "*motion in her voice*" (**P15**). The three participants that correctly felt something was amiss based their discernment on the fact that "*intensity and pauses between words are somewhat flat*" (**P16**). The varying pitch was again the most cited cue of deceptiveness, next to the distinct appearance of inflection, for clip **E6** because "*she sounds heavier on the beginning of one word and the end of the next, she matches it with the kind of emotion adequate for the meanings of the word*" (**P16**). The minority of the participants correctly felt clip **E6** because the speaker "*doesn't take the time to pause, mid-sentence*" (**P13**). Half the sample that incorrectly discerned clip **E7** as legitimate based their decision on the fact that the "*The speech had inflection that one doesn't generally hear in robotic voices*" (**P2**). The other half was correctly convinced they heard distortions as the voice "*sounded almost like it had like an echo to it.*" (**P6**). All but one of the participants were deceived into believing that the clip **E8** was legitimate by the immediate impression that the speaker "*had a rising pitch at the end*" (**P15**).

4.2.3 Discernment Performance. The Appendix [35] breaks down the discernment cues per the assessment outcome and per the set of stimuli. For the *linguistic expression* of the ASVspoof set of stimuli, incorrectly discerning a legitimate audio clip as a fake was mostly based on the absence of emotion or fluency in the voice of the speaker, while the incorrect discernment of a fake audio clip as a legitimate was based on the presence of inflection and human-sounding voice. For the *acoustic expression*, what made our participants discern the legitimate clips as fake was the pitch's flatness and the speaker's lack of intensity. Analogously, convincing speech of the deepfakes was potent enough when it had variation in the pitch and intensity for the participants to discern them as legitimate. The *recording quality* was less important as a cue for our participants, though the perception of background noise (or absence of it) made them think the legitimate clips were fake and vice versa. The perception of a robotic or a synthesized *recording identity* – driven perhaps by the exposure to TTS audio that our blind or low vision participants depend on for access [16] – was the reason they mistakenly took the legitimate clips as fake. Equally, the perceived naturalness in the speech, such as a conversational delivery or making a statement, was a cue that mistakenly took the fake audio clips as legitimate.

The same cues of incorrectly discerning a legitimate audio clip as a fake (absence of emotion or fluency) and vice versa (presence of inflection and human-sounding voice) were ventured by our participants relative to the *linguistic expression* of the ElevenLabs set of stimuli. Relative to *acoustic expression*, what made our participants discern the legitimate clips as fake was the flatness of the pitch and the perceived imperfections in the computer generation of the voice. What passed as legitimate but was fake was the same generation effect as in the ASVspoof2021 stimuli – a noticeable variation in the pitch and the intensity of the speech delivery. Here too, the *recording quality* was less important and mostly driven towards the perception of recording distortions (or the absence of them) that informed our participants' incorrect assessments. Again, being regularly exposed to a synthesized TTS voice was a sufficient experience for our participants to use it as a discriminator against natural speech, even in deceptive circumstances.

When making correct decisions about the legitimate (*bona fide*) ASVspoof2021 stimuli, our participants mostly relied on the presence of natural articulation and inflection in the speaker's voice. The correct discernment of fake clips, here, was equally predicated on the absence of natural articulation (i.e., "robotic" or "synthesized" impressions of speech), either too flat or too perfect pitch in the voice, or presence of recording distortions (our participants felt that these are the result of the computer generation of the voice such as echo, clippings, or reverberations). The correct discernment of the legitimate ElevenLabs clips was cued by the fluency of the speakers, the distinct variation in their voice, and the perceived natural delivery that often expressed emotion. The cues that lead to a correct determination of the deepfakes as computer generated, here, mostly revolved around the too perfect (or absence of natural) fluency and pitch and perfect sounding recording.

4.3 RQ3: Discernment Comparison between Blind and Low Vision Individuals

4.3.1 Detection Performance. The accuracy performance, broken down per the participants' visual disorder self-identification to "blind" and "low vision," (Table 2), is given in Table 5. The part of the sample that self-identified as "blind" achieved a 55% accuracy while the participants that self-identified as "low vision" achieved a 61% accuracy, over the entire set of study stimuli. Broken down per segment of stimuli, both the "blind" and "low vision" participants performed comparably for the ASVspoof2021 stimuli – 63% versus 64%, respectively – while the "low vision participants performed better with the ElevenLabs stimuli – 59% versus 48%, respectively. Two of the "blind" participants (**P12** and **P15**) and two of the "low vision" participants (**P10** and **P13**) achieved the highest accuracy of 69% (5 total misidentifications) across all the stimuli. The lowest accuracy in the "blind" part of the sample was 44% (9 misidentifications, both **P1** and **P7**). The lowest accuracy in the "low vision" part of the sample was 50%, achieved by **P5**.

Relative to the accuracy per legitimate audio clip, we noticed that the participants from the "low vision" part of the sample performed much better with the ElevenLabs part of the stimuli, achieving 100% accuracy for three out of four clips (**E1**, **E3**, and **E4**). The "blind" counterparts were not that much off in the performance, achieving 88% for clips **E1** and **E4**, and 75% for clip **E3**. Both groups of participants dropped the performance to 38% accuracy for clip **E2**. For the legitimate part of the ASVspoof stimuli, the "low vision" participants were slightly better compared to the "blind" ones (59% versus 50%). The worst accuracy in the prior group was 50% for

Table 5: RQ3: Accuracy Performance (Blind vs Low Vision): Legitimate vs. Fake

	ASVspoof2021 – "Blind" Accuracy 63%								ElevenLabs – "Blind" Accuracy 48%								
P	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**Acc**
1	Legit	Fake	Fake	Fake	Legit	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Legit	Legit	44%
2	Legit	Legit	Legit	Fake	Legit	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Legit	Legit	Legit	Legit	56%
3	Fake	Legit	Legit	Fake	Fake	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Legit	Legit	56%
4	Fake	Legit	Legit	Legit	Fake	Fake	Fake	Legit	Fake	Legit	Fake	Legit	Legit	Legit	Fake	Legit	56%
7	Fake	Fake	Legit	Fake	Legit	Fake	Fake	Legit	Legit	Fake	Legit	Fake	Legit	Fake	Legit	Fake	44%
12	Legit	Fake	Legit	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Legit	Fake	Fake	Legit	69%
14	Legit	Fake	Legit	Fake	Fake	Fake	Fake	Legit	Legit	Fake	Fake	Legit	Legit	Legit	Fake	Legit	50%
15	Fake	Legit	Legit	Fake	Fake	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Fake	Fake	Legit	69%
Acc	50%	50%	88%	13%	63%	100%	75%	63%	88%	38%	75%	88%	0%	38%	50%	13%	
P	**ASVspoof2021 – "Low Vision" Accuracy 64%**								**ElevenLabs – "Low Vision" Accuracy 59%**								
5	Fake	Fake	Legit	Legit	Legit	Fake	Legit	Legit	Legit	Fake	Legit	Legit	Fake	Fake	Legit	Legit	50%
6	Fake	Legit	Legit	Fake	Fake	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Fake	Legit	63%
8	Legit	Fake	Legit	Fake	Fake	Fake	Legit	Legit	Legit	Fake	Legit	Legit	Fake	Fake	Fake	Legit	63%
9	Legit	Legit	Fake	Fake	Fake	Fake	Fake	Legit	Legit	Legit	Legit	Legit	Legit	Legit	Fake	Legit	63%
10	Legit	Legit	Legit	Legit	Fake	Fake	Legit	Legit	Legit	Legit	Legit	Legit	Legit	Fake	Legit	Legit	69%
11	Legit	Fake	Legit	Legit	Fake	Fake	Fake	Fake	Legit	Fake	Legit	Legit	Legit	Legit	Legit	Legit	63%
13	Legit	Fake	Fake	Legit	Fake	Fake	Fake	Legit	Legit	Legit	Legit	Legit	Legit	Fake	Legit	Legit	69%
16	Legit	Legit	Fake	Fake	Legit	Fake	Fake	Legit	Legit	Fake	Legit	Legit	Fake	Legit	Fake	Legit	56%
Acc	75%	50%	63%	50%	75%	100%	63%	25%	100%	38%	100%	100%	38%	50%	50%	0%	
	Color Scheme: Light Violet – Legitimate, Light Yellow - Fake, Gradient Gray - Accuracy **Color Gradient**: Proportional to the accuracy percentage																

the clips **A2** and **A4**, while for the latter group was for 13% for clip **A4**. Overall, the participants in the "blind" group showed better assessment performance than the "low vision" group for clips **A3**, **A7**, **A8**, and **E8** (4 clips), the "low vision" group performed better for clips **A1**, **A4**, **A5**, **E1**, **E3**, **E4**, **E5**, **E6** (8 clips), and both groups were tied in the performance for **A2**, **E2**, **A6**, and **E7** (4 clips).

Per deepfake audio clip, the participants self-identifying as "blind" performed slightly better with the ASVspoof2021 part of the stimuli compared to those that self-identified as "low vision" (75% versus 69%). The critical difference was the clip **A8** were the majority of the "blind" participants correctly felt that "*the speech was too monotone*" to be real **P2**. The detection accuracy performance of our sample noticeably dropped for the audio clips generated using the ElevenLabs algorithm, with the "low vision" participants performing with an overall accuracy of 38% compared to 25% among the group of "blind" participants. Noticeably, the "low vision" group of participants achieved a 0% accuracy for clip **E8** while the "blind" group of participants did the same for clip **E5**. For clip **E8**, the "low vision" participants were tricked by the feeling that the voice was "*three dimensional*" (**P8**). For the clip **E5** the "blind" participant felt there was a "*natural variability in the voice*" **P7**.

The overall performance comparison between the participants in the "blind" and the "low vision" groups is given in Table 6. Precision-wise, "blind" participants performed better for the ASVspoof2021 stimuli than the ElevenLabs ones (60% vs 47%, respectively), while the "low vision" participants showed an opposite performance (62% vs 64%, respectively). Recall-wise, both groups showed a significant drop for the ElevenLabs stimuli – (25% "blind" vs 34% "low vision," respectively). Relative to the F_1 score, the "blind" participants were less reliable compared to their "low vision" counterparts when they assessed the ElevenLabs audio deepfakes (33% vs 46%, respectively) but were more reliable when they assessed the ASVspoof2021 deepfakes (67% vs 64%, respectively). "Low vision" participants were slightly better in correctly selecting a legitimate audio when assessing both the ASVspoof2021 clips (Specificity: 59% vs 50%) and the ElevenLabs clips (Specificity: 72% vs 84%).

Table 6: RQ3: Performance - "Blind" vs "Low Vision"

Metric	ASVSpoof2021	ElevenLabs	Overall
Self-identified as "Blind"			
Accuracy	0.62	0.48	0.55
Precision	0.60	0.47	0.56
Recall	0.75	0.25	0.50
F_1 **Score**	0.67	0.33	0.53
Specificity	0.50	0.72	0.61
Self-identified as "Low Vision"			
Accuracy	0.62	0.59	0.61
Precision	0.62	0.69	0.64
Recall	0.66	0.34	0.50
F_1 **Score**	0.64	0.46	0.56
Specificity	0.59	0.84	0.72

4.3.2 Detection Approach. We broke down the thematic analysis of the decision-making approach shared by our participants based on their visual disorder self-identification in Appendix [35]. For the ASVspoof2021 stimuli, the participants in the "blind" part of the sample mostly relied on a combination of *inflection and imperfections in the voice of the speaker* to guide their decisions about the provenance of each clip. Their "low vision" counterparts for mostly focused on the *pitch, enunciation, and delivery of emotion* as discriminators between legitimate and fake for each clip. For the

ElevenLabs stimuli, the "blind" participants, in addition, considered the *intensity in the speech delivery* to be a factor that discriminates between a real speaker and a speech they regularly encounter as synthesized by screen readers. The "low vision" participants, here, considered the *fluency in expression and articulation* in addition when deciding if a clip was real or fake. Seen from a perspective of the assistive technolog(ies) used by our participants (Table 2), the overall sense whether a speech sounds robotic, synthesized, or natural was more pronounced in the decisions among the participants in the "blind" group.

When accessing the **A1** clip, the participants from the "blind" group who correctly identified it as legitimate heard "*authentic breath and inflection*" **P1**, similar to the assessment of their "low vision" counterpart **P11**. The "blind" participants that accessed the **A1** incorrectly as a fake, such as **P3**, felt that the clip was very "*robotic sounding*," which came close to the feeling of speech delivery "*without much intensity and normal fluency*" assessment by *P6* from the "low vision" group. The correct assessments of the **A2** clip among the "low vision" participants revolved around "*pitch and enunciation* (**P9**), while among the "blind" participants around being "*naturally conversational*" (**P4**). The incorrect assessments felt the voice was "*flat*" (**P8**) ("low vision") and "*emotionless*" (**P14**) ("blind"). Our "blind" participants correctly identified the **A3** clip as legitimate in all but one of the cases where **P1** felt the "*voice sounded more synthesized.*" In the "low vision" group the correct detection was driven by "*natural inflection quality*" (**P6**) and the incorrect detection by the sense of "*robotic quality*" (**P16**).

The participants in the "blind" group were incorrect in all but one of the cases where participant **P4** correctly identified clip **A4** as legitimate because "*it sounded very matter-of-fact.*" The "low vision" participants were split in their detection accuracy with the correct ones stating that the clip was "*pick up on breathing tones at the end of it*" (**P11**) while the incorrect ones felt it was "*flat*" (**P16**). Both the "blind" and "low vision" participants correctly spotted the **A5** clip as a deepfake because of its robotic feel (**P15** and **P8**, respectively), but some of them were tricked by the "*humanistic vibe*" of the voice (**P7** and **P5**, respectively). The **A6** clip was the only one that all the participants correctly said it was fake, regardless of where they fell on the spectrum of their visual disorder. Clips **A7** and **A8** tipped both "blind" and "low vision" participants to correctly identify them as deepfakes because both sounded "*very patterned*" (**P13**, and **P1**, respectively). Conversely, clips **A7** and **A8** tricked both "blind" and "low vision" participants into identify them as legitimate because the words in both "*flow very well together*" (**P8** and **P14**, respectively).

Both groups performed similar for the clips **E1** and **E4**, where only one participant in the "blind" group and none in the "low vision" group took them as fake because "*the pauses (speaker's fluency) are just wrong in between the words*" (**P4** and **P7**). For the **E2** clip, the participants in the "blind" group correctly pointed out to the "*shape to [the speakers] speech*" (**P2**) and the ones in the "low vision" group correctly felt the speaker had a natural "*human inflection*" (**P13**). Those in both groups that were tricked to believe this clip was a deepfake felt the speaker too closely resembled "*synthesized speech*" they usually encounter as an access medium though their screen readers (**P8** and **P14**, respectively). While all the "low vision" participants were correct that the clip **E3** was legitimate, couple of "blind" participants felt otherwise because they sensed an "*absence of intensity in the delivery*" **P14**.

All the participants in the "blind" group correctly detected that the clip **E5** was fake, while three participants in the "low vision" group were tricked to believe otherwise because it sounded "*pretty real based on the vocal tones*" (**P9**). The "*tone change*" was also the reason why all but three participants in the "blind" group and half in the "low vision group" fell for the deception of clip **E6**. The 50/50 split in the assessment for both groups was the case for clip **E7** where the presence or absence of "*inflection and speech distortions*" (**P14** and **P10**, respectively) were the reasons why participants felt one way or the other. The "low vision" participants were all wrong in their assessment of **E8** as a legitimate, tricked by how "*real the tone sounds*" (**P11**) , and only one participant in the "blind" group called the deception by the "*had a rising pitch at the end*" (**P15**).

4.3.3 Discernment Performance. The Appendix [35] also breaks down the discernment performance per cue used, per the set of stimuli, and per visual disorder self-identification. Participants that self-identified as "blind" incorrectly discerned a legitimate audio clip as a fake from the ASVSpoof2021 when they relied on cues such as absence of inflection and emotion as well as the feeling of robotic or synthesized-sounding voice. Participants in the "blind" group made an incorrect discernment of a fake audio clip as a legitimate when they rested their decision on the presence of inflection and emotion and feeling of natural-sounding voice. The similar cues were also used for the ElevenLabs stimuli, with a noticeable over-reliance on the presence of pitch variations as a hallmark of a legitimate clip when in fact the audio was deepfake. Our "blind" participants were correct about the legitimate clips when focused on the imperfections in the recordings and about the fake clips when focused on the articulation and speech synthetization .

Participants that self-identified as "low vision" incorrectly discerned a legitimate audio clip as a fake from the ASVSpoof2021 when they relied mostly on the fluency, intensity, and the absence of variation in the speaker's pitch. On the other hand, they incorrectly labeled a fake audio clip as a legitimate when they were convinced the speaker's articulation sounded natural. Similarly, for the ElevenLabs stimuli, the perceived lack of naturalness and emotion on the speaker side were the incorrect cues that a legitimate audio was fake, while fluency of the speaker and the variations in the pitch were sufficiently convincing cues to trick them into believing that a deepfake audio was legitimate. Our "low vision" participants made correct decisions about the legitimate clips when they considered the articulation and the naturalness of the speech in the context of the imperfections in the speaker's voice and made correct decisions about the fake when they noticed unnatural pitch, robotic recording identity, and too perfect of an expressive fluency.

4.4 RQ4: Perspectives on Audio Deepfakes

After our participants completed the assessment and discernment tasks, we shared their performance and asked them to comment on it. Given that none of our participants performed well relative to correctly discerning the legitimate speech from the deepfake one, the general impression in our sample was that the audio deepfakes are "*little scary, given how impressively a technology could impersonate a speaker*" (**P2**) For individuals who depend on "synthesized

speech," our participants pointed out that "*a bit of context for each clip would have helped, perhaps a bit*" (**P4**) but the discernment task would have "*remained unsurprisingly challenging*" (**P6**). What perhaps threw most of them was how "*natural the speech sounded*" (**P7**) as they felt they are "*pretty good in sensing an authentic speaker*" (**P4**), given their everyday exposure to synthesized voice.

As the analysis above shows, the blind and low vision participants in our sample heavily rely on speech naturalness, including conveying emotion and hearing breathing. So the ability of technology to fake this aspect was seen as a "*dangerous watershed moment*" (**P13**) because now "*a sudden uncertainty appears in hearing not just any audio online but also second-guessing the screen reader, audiobooks, narration, and any auditory media*" (**P9**) they consume and depend on. Participants, like **P3** made an immediate connection with the ability to generate "*synthetic text with technologies like ChatGPT*" and use that as an opportunity to "*create a multidimensional deceptive environment where even the context or metadata about an audio might fake and blind people have no way of knowing it*" Regardless and quite inspiring, participants in our sample did not felt discouraged by the hyphened and undue risk of the all-round deceptiveness. Instead, many of them stated that the participation in the study "*inspire them to look more about these audio deepfakes and keep track of them*" (**P8**) and "*pay more attention to the quality of voices*" (**P16**) they hear every day.

Next, we asked our participants to share their perspectives on aspects or scenarios where an audio deepfake could be used to deceive blind or low vision individuals. Five main scenarios emerged as of particular concern for our participants: (i) political disinformation; (ii) phone scams and voice phishing; (iii) fake evidence in legal proceedings or accusations; (iv) voice authentication; and (v) tricking voice assistants. When it comes to political disinformation, our participants saw a broader opportunity for bad actors to particularly target blind individuals with "*sensationalist audio deepfakes as that became a trend in the US politics*" (**P2**) because it "*hard to discriminate between a flood of convincing public speeches, even if only parts of them are fake* (**P1**). One of the participants, **P11**, drove the point about disinformation proliferation by stating that "*what memes are usually for sighted people, audio deepfakes are a perfect counterpart for blind or low vision people*" alluding to the most exploited format of misinformation encountered online.

Our participants saw a double exposure to the risk of convincing speech as "*now it seems plausible to create a human-like screen reader and feed it with a lot of textual fake news or misleading political information*" (**P6**). This misleading information was obvious to our participants that could also be used in the usual phone scams that they are constantly targeted with, but they saw an opportunity for a scammer to "*use multiple voices, even celebrity-sounding ones or clone one's from people's Instagram or TikTok, to make an approach both by calling them and leaving a voice message*" so they lure them step-by-step (**P13**). Relative to voice phishing, our participants envisioned a plausible pretexting where one could "*clone my boss's voice or my supervisor's voice, and tell me to like turn over certain work information*" (**P1**) akin to the actual attack attempted to the LastPass employee. Interestingly, our participants envisioned a scenario where audio deepfakes are used to "*create a fake evidence impersonating someone and used that in criminal proceedings*" (**P4**).

Suspecting a foul play on the law enforcement's side, for example, participant **P3** saw an opportunity for "*planting an evidence to get someone wrongly convicted*." An important scenario that got our participants concerned is the appeal of voice passwords to blind or low vision individuals, which now, with the appearance of an easy audio deepfake technology, are under "threat to easily misused, even in a two-factor authentication is enabled" (**P5**). The obvious concern was their "*bank accounts being compromised*" before being too late, as it "*takes quite a lot of time for a blind person to restore their passwords and bank balance*" (**P3**). Voice assistants with the ability to control one's physical security, for our blind and low vision participants, were of particular concern. Participant **P10** saw an opportunity where one could use a "*loud enough fake recording of their voice to target their Amazon Alexa to simply open a door*."

5 Discussion

Our results show that the deceptiveness of the deepfakes as real speech and sense of fake quality of otherwise legitimate speech among our participants was non-negligible. While the overall discernment accuracy was 59%, our results show a slight advantage in discernment of the participants that self-identified as "low vision" (accuracy of 62%) compared to the ones that self-identified as "blind" (accuracy of 55%). Broken down per type of deepfake algorithm, the performance analysis reveals an interesting aspect of the ASVspoof part where the blind and low vision participants performed considerably better when detecting a TTS-generated speech (algorithms A9, A7, A15) than the voice-cloned one (A19). The difference is even more noticeable when the voice cloning is done by the ElevenLabs algorithm, where our participants' precision dropped to a concerning 29.7% of correct discernment. Controlling for the self-identified visual disorder, the participants in the "blind" group demonstrated a better discernment performance when detecting the TTS-generated speech with algorithm A15 and the voice-cloned one with algorithm A19, while the "low vision" participants were better in the case of A7. Both groups were tied in the performance for the TTS-generated speech with algorithm A9.

The experience and nuanced familiarity with TTS audio, however advantageous, is perhaps not sufficient for fending off the general deceptiveness of the audio clips, our qualitative results show. Our participants were predominantely looking for inflection, emotion, pitch variation, fluency, and an intense delivery of the speech. More specifically, the participants in the "blind" part of the sample mostly relied on a combination of inflection and imperfections in the voice of the speaker (ASVspoof2021 stimuli) and also considered the intensity in the speech delivery as an authenticity factor (ElevenLabs stimuli). The participants in the "low vision" used the speaker's pitch, enunciation, and delivery of emotion to discriminate between legitimate and fake speech (ASVspoof2021 stimuli), and also considered the fluency in expression and articulation of the speaker when needed (ElevenLabs stimuli).

The impression of a natural prosody revolved around a clear distinction of a *human voice* that differs from either a robotic, cloned voice or a TTS synthesized one simply because computers "*don't make breathing noises and make predictable, rhythmic pauses*" (**P3**). In other words, what our blind and low vision participants expected as a legitimate voice was one where the speaker clearly inflects

words with emotion and varies the pitch in the attempt to convey a sincere signature that they – in the words of participant **P11** "*speak the sentences as [she] mean it.*" That the deepfake audio we used in our study, especially the ElevenLabs stimuli, were able to mimic these convincing traits of the *human voice* was seen as a warning sign that "*time has come where blind people suddenly don't know who's voice to trust on the Internet anymore* (**P6**).

5.1 Audio Deepfakes and Humans Who Are Blind or Low Vision: Contributions

Relative to the current 'humans versus audio deepfakes' effort, we add the following contributions. We expand the assessment of human ability beyond the measure of accuracy that was central to the work in [11, 16, 25, 27] to include precision and recall to better understand how humans fare particularly to deepfake deception. We also added F_1 and specificity to better understand their ability to remain true when judging legitimate audio (Tables 4 and 6). We used English speaking participants and we didn't test the expanded assessment on individuals who are low vision or blind that speak other languages like the studies with sighted individuals did, but we added an extensive quantitative component that offered an in-depth analysis of the decision-making process behind the discernment. In addition to comparing both the qualitative and quantitative aspects of the discernment between the participants that self-identified as "blind" and the ones that self-identified as "low vision," we also collected their perspectives on how audio deepfakes could affect or are already affecting their real lives.

The only study that involved participants that self-identified as "blind" [16] did collect the qualitative component but it was only limited to the decision-making behind the discernment and it didn't provide any broader perspectives on how people that rely on audio as a medium as well as an affordance think about audio deepfakes. Relative to this particular study, we addressed the limitation of recruiting and comparing the discernment performance of not just blind people that rely on audio as an affordance but also people who self-identify as "low vision" and use assistive technologies such as screen readers too. Unlike the approach in [16], we create a balanced set of stimuli (8 legitimate and 8 fake) that were curated from the AVSpoof2021 dataset to sound realistic without the need for a context and we pilot tested it to ensure that our participants would not experience a fatigue during the quantitative part of the study.

While our selection was on part with the number of stimuli and curation policies used in the studies with the sighted individuals, we paid a specific attention to ensure the stimuli are selected from the most recent public datasets available but also to create deepfakes with the latest available synthetic-voice cloning technology – a methodological aspect of creating realistic study settings that was absent from all the prior work of testing humans versus audio deepfakes. Considering that humans usually demonstrate an accuracy between 50% and 60% on average (Table 1), the performance of "blind" individuals (accuracy of 55%, precision of 56%) and "low vision" (accuracy of 61%, precision of 64%) individuals in our study comes on par with this performance too. Compared specifically to the accuracy of 59% demonstrated by the individuals that self-identified as "blind" in [16], our results show a comparable performance where the negative difference of 4% could be attributed to the particularly deceptive subset of ElevenLabs stimuli. Controlling for the nature of the deepfake (TTS or VC), we noticed a similar, albeit anecdotal trend, where both the "blind" and "low vision" individuals were better at detecting TTS generated deepfakes than VC ones. In addition, we found than the "blind" participants were slightly better than their "low vision" counterparts when assessing VC clips, but were outperformed for the TTS generated ones.

5.2 Resilience to Audio Deepfakes

The established deceptiveness of the audio deepfakes, reinforced by the overall findings of our study, would hardly, in and of itself, help blind and low vision individuals improve their detection and discernment abilities in the future. A natural response to deception in a usable security context would be to (i) raise the awareness about audio deepfakes or (ii) warn about an immediate encounter with a potentially audio deepfake. Raising awareness – be that through *facts and advice*, *games*, or *simulations* – though useful, might lag behind the pace of technology with which a convincing, yet synthesized speech is generated. Usually in these instances individuals are warned about a potential deception through synthetic media, but such warnings exist predominately for video content or images.

A development of such warnings about synthetic audio would be beneficial for blind or low vision users as they predominately rely on this type of media, though their application is predicated on an automated detection itself. A recent trend in response to audio deepfake deception is the use of watermarking, a technique where a detection clue is included in the generated speech to mark it as a "deepfake" in a way that aids detection by another machine, but remains transparent to a human listener [18]. This detection clue is essentially a 128-bit message that is stored as a watermark in the output sound file and could be easily implemented by synthetic voice generators like ElevenLabs through open-source tools such as Audiowmark [1]. Obviously, this approach is driven to work from a machine detection perspective, not from a human perspective.

But the reliance on accessible technology, such as screen readers for individuals that are low vision or blind, in this case, might offer an advantage in generating warnings, if the watermarks themselves are made with accessibility in mind. The screen readers themselves usually only read textual input, but per the WEB-ARIA guidelines [38], accessible media – including audio – should usually be accompanied with a textual information such as captions, transcript, or an audio description. Part of this information could be used to convey the presence of the watermark and therefore can be used to warn a blind or a low vision individual about the provenance of the audio they are about to hear. For example, when one navigates to an audio element that has a play button, the screen reader could access the alternative text element attached to the audio file and narrate it after the audio is played back, as shown in Appendix [35].

Solutions like these would work if the one who generates the audio is willing to include a watermark in the first place. If a deepfake is generated without a watermark, and the LastPass example suggests that this is desirable from an adversarial deception perspective [22], there is little that blind and low vision individuals

[1] https://github.com/swesterfeld/audiowmark

could do. If a non-watermarked deepfake audio is sent through a direct message, this population has no other option but to rely on the cues for detection they also employed when participating in this study. If the non-watermarked deepfake audio is shared on social media, then, platforms have the opportunity to add the aforementioned warning, even if it's not part of the audio description, but part of the interface, as they already do with warning covers that warn users about a synthetic video or images. Of course, these warnings need to be accessible through a screen reader so such a capability is yet to be developed where a screen reader user could navigate to an element on the social media platform that verbalizes a warning that a post might contain synthetic audio (or video) [33].

5.3 Extended Deception of Audio Deepfakes

The (anticipated) resilience efforts, like in the case of phishing, are usually based on the *known deceptiveness* of audio deepfakes at the period when they are developed. But deepfake technology rapidly evolves, and with that, the *unknown deceptiveness* always leaves open avenues for harm to blind or low vision individuals. One such avenue, of course, is improving voice cloning technologies to account for the cues used to correctly spot an audio fake. An adversary could well use the results of our paper or other related papers [16, 27] and simply target voice cloning that makes better inflection, delivers emotion, spaces words and pauses, and even makes the breathing noise or any vocal tract imperfections incorporated in a recording. Another equally dangerous avenue is the use of freely available tools such as ChatGPT to generate text that could be translated into speech and used for deceptive purposes.

Joint use of a generative text and audio – especially for a population that critically relies on text-to-speech assistance – though not yet applied or used, to our knowledge, could be a damning prospect, as our participant **P3** pointed out. One could think of communicating a pretext using ChatGPT-generated prompts and then following up with a voice call to make a convincing reason for a blind or low vision individual to either change a password, install software, or send money. Even that might not be needed, given that ChatGPT [2] itself allows for a spoken back-and-forth conversation – it is sufficient for an adversary to simply use this (or similar) features against a targeted blind or low vision individual. We do not have a ready answer on how to approach this particular type of deception, but we nonetheless advocate for proactive anticipation of such possibilities of extended harm to blind or low vision individuals.

5.4 Level of Visual Perception as a Dimension

Our study recruited 16 participants who identified as being a part of the blind and low vision community, having at least one eye disorder that classifies them as being *legally blind*. While the participants' self-identified as "blind" or "low vision", we recognize that this information alone cannot be reliably used to discern each participant's distinct level of visual perception (e.g., totally blind with no perception of any lights or shapes, blind with the ability to perceive lights only, low vision with the ability to read 20 pt. font on a 27-inch monitor at a brightness of 300 nits, etc.) This distinction limits nuanced observations based on what we used as a categorical variable in our study to perform initial comparisons relative to whether one falls on the visual disability spectrum.

While one could argue that the difference between these various levels of visual perception may be insignificant or completely immeasurable, we could not definitively say without any further extensive testing. Someone who is familiar with the intricacies of the visual disorders that our participants identified with may be able to piece together an idea of where that individual participant falls on the spectrum, however, they would require more information to establish a holistic image of that person's level of visual perception. In the future, our research will strive to contextualize the exposure to deepfake audio (and video) relative to a more nuanced take of this demographic. In addition to the visual diagnosis variable, we are curious to learn more about how our future blind and low vision participants describe, in their own words, what their level of visual perception is, in what capacity they rely on a screen reader, screen magnifier, or both, and how other variables impact their ability to see (e.g., font size, screen size, environment lighting, contrast, etc.). We believe that by doing this, we might uncover a deeper insight into the nature of the deepfake deceptiveness, and with that, help the resilience effort for blind and low vision individuals.

5.5 Limitations

The exposure to deceptive audio clips, however short and in strictly controlled settings, nonetheless imposes several limitations pertaining to our study. A limitation comes from the sample size that prevents the generalization of the results to the entire population of blind and low vision individuals. We were limited to selecting study stimuli from ASVspoof2021, the VCTK corpus, and using ElevenLabs as a voice cloning technology. Other stimuli from the same or other datasets, textual or speech corpora, or voice cloning technologies might cause blind or low vision individuals to perform and behave in a way that differs from the findings in our study. Another limitation is that we sampled people from the US and used sentences spoken in a US English accent. Other languages as well as other accents might yield different results than ours.

We were limited to the audio deepfake technology available up to the first quarter of 2024. Future voice cloning technologies, TTS manipulations, or generative algorithms might render a level of audio deepfake deceptiveness different from the one reported in this study. The current version of assistive technologies could pose yet another limitation as any new features (e.g., realistic voices) might transform how blind or low vision users' approach and discern audio deepfakes and, with that, affect the overall findings. Though we left our participants sufficient time and support to engage with each of the audio clips, this might have not been sufficient for them to formulate a more informed expression about their overall assessment and ultimate decision about it. Despite all these limitations, our study nonetheless provides rich accounts of blind and low vision individuals' first account experiences with audio deepfakes.

6 Conclusion

In this paper, we worked with 16 blind and low vision individuals to address the knowledge gap relative to how well they could discern a legitimate audio from a deepfake audio and learn their perspectives about the deceptiveness of convincing, but computer-generated

[2] https://openai.com/blog/chatgpt-can-now-see-hear-and-speak

speech. Our study reveals that blind and low vision individuals do well discerning if the speech is generated through TTS means, but the performance wanes in cases of advanced voice cloning. The lack of human sounding prosody such as inflection and emotion, paired with the absence of traits such as human breathing were sufficient cues for a blind or low vision individual to discern an audio as a fake. This approach did not always yield accurate results and our participants saw a potential for using audio deepfake in practical deception such as misinformation, voice phishing, scamming, and voice impersonation. To address these deception possibilities, we discuss several usable security approaches in helping blind and low vision individuals build resilience against audio deepfakes.

References

[1] Domna Bilika, Nikoletta Michopoulou, Efthimios Alepis, and Constantinos Patsakis. 2024. Hello me, meet the real me: Voice synthesis attacks on voice assistants. *Computers & Security* 137 (2024), 103617.

[2] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. 2022. Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 2691–2708. https://www.usenix.org/conference/usenixsecurity22/presentation/blue

[3] Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro. 2021. Synthetic speech detection through short-term and long-term prediction traces. *EURASIP Journal on Information Security* 2021, 1 (2021), 2.

[4] Robert Chesney and Danielle Citron. 2019. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff.* 98 (2019), 147.

[5] Cybersecurity and Infrastrucure Security Agency (CISA). 2021. Avoiding Social Engineering and Phishing Attacks. https://www.cisa.gov/news-events/news/avoiding-social-engineering-and-phishing-attacks.

[6] Adrienne de Ruiter. 2021. The Distinct Wrong of Deepfakes. *Philosophy & Technology* 34, 4 (2021), 1311–1332.

[7] Eleven Labs. 2023. Generative Voice AI. https://elevenlabs.io.

[8] Federal Communication Commission. 2024. Chatbots, deepfakes, and voice clones: AI deception for sale. https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale

[9] Federal Communication Commission. 2024. Implications of Artificial Intelligence Technologies on Protecting Consumers from Unwanted Robocalls and Robotexts. https://docs.fcc.gov/public/attachments/FCC-24-17A1.pdf

[10] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. John Wiley & Sons.

[11] J. Frank, F. Herbert, J. Ricker, L. Schönherr, T. Eisenhofer, A. Fischer, M. Dürmuth, and T. Holz. 2024. A Representative Study on Human Detection of Artificially Generated Media Across Countries. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 162–162. https://doi.org/10.1109/SP54263.2024.00159

[12] Joel Frank and Lea Schönherr. 2021. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813* (2021).

[13] Elaine Gerber. 2002. Surfing by ear: Usability concerns of computer users who are blind or visually impaired. *Access World* 3, 1 (2002), 38–43.

[14] Ben Gold, Nelson Morgan, and Dan Ellis. 2011. *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons.

[15] Matthew Groh, Aruna Sankaranarayanan, and Rosalind Picard. 2022. Human detection of political deepfakes across transcripts, audio, and video. *arXiv preprint arXiv:2202.12883* (2022).

[16] Chaeeun Han, Prasenjit Mitra, , and Syed Masum Billah. 2024. Uncovering Human Traits in Determining Real and Spoofed Audio: Insights from Blind and Sighted Individuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (Honolulu, HI, USA). Association for Computing Machinery, New York, NY, USA, 14 pages. https://doi.org/10.1145/3613904.3642817

[17] Simon Hayhoe and Azizah Rajab. 2000. Ethical considerations of conducting ethnographic research in visually impaired communities. In *European Conference on Educational Research*.

[18] Lauri Juvela and Xin Wang. 2024. Collaborative Watermarking for Adversarial Speech Synthesis. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 11231–11235. https://doi.org/10.1109/ICASSP48485.2024.10448134

[19] A. Kassis and U. Hengartner. 2023. Breaking Security-Critical Voice Authentication. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 951–968. https://doi.org/10.1109/SP46215.2023.10179374

[20] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain De Cheveigne. 1999. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech communication* 27, 3-4 (1999), 187–207.

[21] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. 2021. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=TAXFsg6ZaOl

[22] LastPass. 2024. Attempted Audio Deepfake Call Targets LastPass Employee. https://blog.lastpass.com/posts/2024/04/attempted-audio-deepfake-call-targets-lastpass-employee

[23] Elaine Lau and Zachary Peterson. 2023. A Research Framework and Initial Study of Browser Security for the Visually Impaired. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Anaheim, CA, 4679–4696. https://www.usenix.org/conference/usenixsecurity23/presentation/lau

[24] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Hector Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. 2023. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 2507–2522. https://doi.org/10.1109/TASLP.2023.3285283

[25] Kimberly T. Mai, Sergi Bray, Toby Davies, and Lewis D. Griffin. 2023. Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE* 18, 8 (08 2023), e0285333–.

[26] Driss Matrouf, J-F Bonastre, and Corinne Fredouille. 2006. Effect of speech transformation on impostor acceptance. In *2006 IEEE international conference on acoustics speech and signal processing proceedings*, Vol. 1. IEEE, I–I.

[27] Nicolas M Müller, Karla Pizzi, and Jennifer Williams. 2022. Human perception of audio deepfakes. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 85–91.

[28] Margi Murphy, Rachel Metz, and Mark Bergen. 2024. AI Startup ElevenLabs Bans Account Blamed for Biden Audio Deepfake. https://www.bloomberg.com/news/articles/2024-01-26/ai-startup-elevenlabs-bans-account-blamed-for-biden-audio-deepfake

[29] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi H. Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md Sahidullah, Junichi Yamagishi, and Kong Aik Lee. 2021. ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 2 (2021), 252–265. https://doi.org/10.1109/TBIOM.2021.3059479

[30] Michelle O'Reilly and Nicola Parker. 2013. 'Unsatisfactory Saturation': a critical exploration of the notion of saturated sample sizes in qualitative research. *Qualitative Research* 13, 2 (2013), 190–197. https://doi.org/10.1177/1468794112446106

[31] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *ACM Comput. Surv.* 54, 4, Article 81 (may 2021), 43 pages. https://doi.org/10.1145/3386867

[32] Filipo Sharevski, Peter Jachim, Paige Treebridge, Audrey Li, Adam Babin, and Christopher Adadevoh. 2021. Meet Malexa, Alexa's malicious twin: Malware-induced misperception through intelligent voice assistants. *International Journal of Human-Computer Studies* 149 (2021), 102604.

[33] Filipo Sharevski and Aziz Zeidieh. 2023. Designing and Conducting Usability Research on Social Media Misinformation with Low Vision or Blind Users. In *Proceedings of the 16th Cyber Security Experimentation and Test Workshop* (Marina del Rey, CA, USA) *(CSET '23)*. Association for Computing Machinery, New York, NY, USA, 75–81. https://doi.org/10.1145/3607505.3607525

[34] Filipo Sharevski and Aziz Zeidieh. 2024. Assessing Suspicious Emails with Banner Warnings Among Blind and Low-Vision Users in Realistic Settings. In *33st USENIX Security Symposium (USENIX Security 24)*. USENIX Association, Philadelphia, MA, 2691–2708. https://www.usenix.org/conference/usenixsecurity24/presentation/sharevski

[35] Filipo Sharevski, Aziz Zeidieh, Jennifer Vander Loop, and Peter Jachim. 2024. Full Version: Blind and Low-Vision Individuals' Detection of Audio Deepfakes. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security* (Salt Lake City, UT) *(CCS '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3658644.3690305

[36] Verizon. 2023. *Data Breach Investigations Report*. Technical Report. Verizon. Retrieved August 31, 2023 from https://www.verizon.com/business/resources/Tabb/reports/2023-data-breach-investigations-report-dbir.pdf

[37] Elliott Vittoria and Kelly Makena. 2024. The Biden Deepfake Robocall Is Only the Beginning. https://www.wired.com/story/biden-robocall-deepfake-danger/

[38] Web Accessibility Initiative (WAI). 2022. Making Audio and Video Media Accessible. https://www.w3.org/WAI/media/av/.

[39] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). (2019). https://datashare.ed.ac.uk/handle/10283/3443.

[40] Heiga Zen, Yannis Agiomyrgiannakis, Niels Egberts, Fergus Henderson, and Przemysław Szczepaniak. 2016. Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices. *arXiv preprint arXiv:1606.06061* (2016).