Uncovering Human Traits in Determining Real and Spoofed Audio: Insights from Blind and Sighted Individuals

Chaeeun Han
Pennsylvania State University
University Park, PA, USA
cfh5554@psu.edu

Prasenjit Mitra
Pennsylvania State University
University Park, PA, USA
pum10@psu.edu

Syed Masum Billah
Pennsylvania State University
University Park, PA, USA
sbillah@psu.edu

ABSTRACT

This paper explores how blind and sighted individuals perceive real and spoofed audio, highlighting differences and similarities between the groups. Through two studies, we find that both groups focus on specific human traits in audio–such as accents, vocal inflections, breathing patterns, and emotions–to assess audio authenticity. We further reveal that humans, irrespective of visual ability, can still outperform current state-of-the-art machine learning models in discerning audio authenticity; however, the task proves psychologically demanding. Moreover, detection accuracy scores between blind and sighted individuals are comparable, but each group exhibits unique strengths: the sighted group excels at detecting deepfake-generated audio, while the blind group excels at detecting text-to-speech (TTS) generated audio. These findings not only deepen our understanding of machine-manipulated and neural-renderer audio but also have implications for developing countermeasures, such as perceptible watermarks and human-AI collaboration strategies for spoofing detection.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **HCI theory, concepts and models**; • **Security and privacy** → *Social aspects of security and privacy*; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Audio perception; generative AI, neural speech, deep fake audio; sighted, blind, vision impairments; audio, speech, voice, text-to-speech (TTS); bona fide audio, spoofed audio, replay attack; and audio watermarking, and human-AI collaboration.

ACM Reference Format:
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), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3613904.3642817

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642817

1 INTRODUCTION

Humans primarily communicate via speech, which contains unique biometric data characterized by the vocal tract shape, larynx size, and other voice production elements [32, 62]. Accent, intonation, pronunciation, and vocabulary further distinguish individuals [32]. Consequently, speech features have grown vital for biometric authentication, especially in automatic speaker verification (ASV) systems [29, 43, 57, 59]. These systems, renowned for conveniently identifying individuals through voice, are prevalent in telebanking, smart speakers, and call centers. The rise of artificial intelligence (AI), particularly generative AI tools like [18, 53], blurs the line between reality and fabrication [52]. Soon, the average person may struggle to discern truth in the face of AI-generated photos, audio, text, and video [41].

This development poses a danger to all; however, blind individuals face a particular disadvantage because they must determine the authenticity of a video clip based solely on its audio channel, while sighted users can analyze both audio and video channels. For instance, sighted individuals can detect inconsistent lighting, shadows, and reflections in the video or observe unnatural facial movements, expressions, and blinking that do not align with speech [6, 22, 44]. Thus, it is crucial to investigate whether the audio channel alone provides sufficient information for blind individuals to assess the authenticity of an audio clip.

This paper aims to investigate what qualities (if any), when present in the audio, allow blind individuals to perceive the audio as spoken by a real human (hereafter referred to as ***bona fide***) or manipulated by an adversary (***spoofed***, hereafter). Currently, *four* different techniques exist for speech manipulation: *i) impersonation*: the adversary alters their voice to resemble that of the target person; *ii) replay*: the adversary records the target person's voice, presumably surreptitiously, and replays it to fake that person's identity; *iii) speech synthesis*: the adversary generates entirely artificial speech signals using rule-based Text-to-Speech (TTS) engines or learning-based AI techniques (e.g., deep fake audio); and *iv) voice conversion*: the adversary uses a system that converts their natural speech to mimic the speech of the target speaker [36].

The prevalence of impersonation remains somewhat uncertain due to the involvement of trained human vocal actors; however, the remaining three spoofing techniques are technology-driven and thus widespread. Replay is the most common and straightforward technique to implement, followed by speech synthesis using TTS, voice conversion, and more recently, speech synthesis with deep fake [37, 59]. This paper focuses on technology-driven spoofing techniques. Specifically, we pose the following research questions (RQs):

- ***RQ1***: *How do blind individuals determine whether an audio clip is bona fide or spoofed, and what attributes (if any) contribute to their decision-making process?*
- ***RQ2***: *How, if at all, does the decision-making process of blind individuals differ from that of sighted individuals?*

To this end, we employed a mixed-method approach. First, we conducted a one-on-one study with 12 blind participants, during which participants assessed 63 *challenging* audio clips as bona fide or spoofed, and walked us through their decision-making processes. We define challenging audio clips as those that the current state-of-the-art anti-spoofing binary classifiers misclassify with high confidence (i.e., detection accuracy: 0%). Readers can also listen to some of these audio clips from this anonymized YouTube playlist[1]. The exact configuration of each audio clip is available in the video description on YouTube. Second, we conducted an online survey with 96 challenging audio clips, where 60 participants – 30 blind and 30 sighted – rated whether clips are bona fide or spoofed, and provided us with an open-ended response explaining their decisions. Finally, we analyzed the findings from both studies, consolidated the results, and discussed the implications of our results, as well as proposed new design considerations to address the broader, societal impacts of spoofed audio.

Our interview study reveals that blind individuals attend to specific human traits in the audio (Section 4). These traits include the *i) speaker's accents* (e.g., uncommon pronunciation and international accents); *ii) vocal inflection* (e.g., monotone, intonation, emphasis, fluctuation); *iii) the sign of liveliness* (e.g., the presence of breathing and lip movement sounds, as well as the irregular pausing patterns); *iv) the presence of emotion in the audio* (e.g., sadness and boredom are associated with spoofed audio and excitement with bona fide audio); *v) the presence of human errors* (e.g., mistakes, mispronunciations, and use of filler words); and *vi) acoustic properties* (e.g., audio quality, and the presence and position of echo and reverberation).

Our survey study, with twice the participants and data points, corroborated our initial findings. Both blind and sighted participants focus on human traits in audio for authenticity. However, their mental models differed: blind participants excelled at identifying TTS audio (91% vs. 84%), while sighted participants were better at detecting deep fake audio (71% vs. 58%). Both groups found replayed audio challenging, with sighted individuals more accurate at close range and blind individuals at longer distances.

Our findings yield key insights. First, both blind and sighted individuals struggle with audio authenticity, showing less than 65% accuracy in recognizing bona fide audio. This implies cognitive strain. Second, both groups apply social norms to judge audio, attending to emotional and physiological cues. With evolving spoofing techniques, the distinction between human and AI-origin audio will blur, shifting focus toward risk minimization. Third, human judgment surpasses current state-of-the-art (SOTA) models. Thus, short-term prospects lie in developing AI models that isolate human traits from audio, breaking down spoofing countermeasures into digestible components, and combining these for more effective solutions. Lastly, AI watermarking must evolve to be more human-focused, considering perceptible cues that both blind and sighted individuals can readily identify. Our findings thus deepen our understanding of machine-manipulated audio and have implications for developing effective, human-centered strategies for spoofing countermeasures in both the short-term and long-term.

[1] https://www.youtube.com/playlist?list=PLf_Q2-kgSq_Dr44vFUopN2jWpqD1gK-7H

2 BACKGROUND AND RELATED WORK

We overview the literature related to our study, examining audio perception, and the effect of visual loss on auditory perception. We then briefly describe two spoofing countermeasures for detecting bona fide and spoofed audio.

2.1 Perception of Speech in Humans

Converting speech into meaningful words is a complex process. It begins when sound waves reach the inner ear, vibrating the organ of Corti. This vibration prompts hair cells to convert the motion into electrical signals, which are then sent through the auditory nerve to the primary auditory cortex. Here, phonemes—individual sounds that comprise words—are recognized [12]. The signals also travel to Wernicke's area and other brain regions, where words are identified and their associated meanings are retrieved [45].

Various psychophysical models explore how our brain processes sound into words [7]. Some models focus on segmenting sounds into discrete words but struggle with ambiguous word boundaries [14]. Others suggest that the brain evaluates multiple potential word sequences to match the incoming audio [39]. Still, other research suggests that speech processing may occur non-sequentially; future sounds can retroactively influence our interpretation of earlier ones [16]. Despite decades of advancements, the perception of speech remains an active area of research.

2.2 Audio Processing Abilities of Blind Individuals

Blind individuals interact with computers using assistive technologies, such as screen readers, which vocalize onscreen content through machine-generated audio synthesized by Text-to-Speech (TTS) technologies [9]. Consequently, they have extensive experience listening to both synthesized (spoofed by default) audio during computer interaction and authentic human voices in interpersonal communication. Moreover, many blind individuals exhibit significantly higher listening rates than sighted individuals due to the human brain's plasticity [8, 11, 12, 28]. This enhanced listening ability suggests that blind individuals' brains process auditory information differently from sighted individuals [54]. For example, brain scans of blind individuals have revealed that they engage the visual cortex, a major region in the human brain, for various cognitive processes when performing tasks or exposed to stimuli [48, 54, 60]. This contradicts the traditional belief that the visual cortex is exclusively reserved for processing visual stimuli (see Kolarik et al. [33] for a review). Because of their extensive experience in differentiating natural human voices from synthesized audio (generated by assistive technologies), we studied blind individuals in great detail.

2.3 Relationship of Auditory Perception with Vision Loss

The differences in how blind and sighted individuals perceive various acoustic properties in audio are nuanced. For instance, the relationship between the severity of visual loss and changes in auditory abilities remains unclear, both in terms of proportionality and systematic effects. Recent research shows that more severe visual loss correlates with increased auditory judgments of distance and room size [33]. This finding is particularly intriguing given that our study employs audio clips—both bona fide and spoofed—recorded in various acoustic environments, encompassing different room sizes, reverberation times, and distances between the sound source and microphone (see Table 2).

Individuals with severe visual impairment perceive sound as twice as distant and rooms as three times larger than do their sighted counterparts [33]. As visual impairment worsens, accuracy in estimating room size improves, but distance estimates become less accurate for closer sounds and more exaggerated for farther ones. Sighted individuals, in contrast, more accurately estimate distance for closer sounds but falter for more distant ones. Accurate judgments of closer sounds are crucial for rapid motor responses like collision avoidance. Our work extends previous studies by linking the perception of audio realism to the acoustic environment in which the audio originated.

Both blind and sighted individuals understand natural speech more easily than synthesized speech. However, blind individuals outperform their sighted counterparts in comprehending synthesized speech, likely due to their greater use of screen readers [42]. Paradoxically, when identifying automatic speech recognition (ASR) errors in dictated text via text-to-speech (TTS) output, blind individuals catch only 40% of ASR errors [27], which is fewer than the 50% detected by sighted individuals [26]. These differences highlight the significant impact of vision loss on speech perception. As such, we included both sighted and blind participants to understand their perception of authentic and spoofed audio.

2.4 Spoofing Countermeasures: Binary Classifiers and Digital Watermarking

One approach to detect spoofed audio is to design powerful binary classifiers to predict the likelihood of a clip being bona fide or spoofed. These classifiers recently use Deep Neural Network (DNN)-based architecture [35, 36, 50, 58, 63–65]. They extract acoustic properties of the input audio by using Fourier transformations, inverse Fourier transformations, and Cepstral analysis [10], such as Mel Frequency Cepstral Coefficients (MFCC) [17] and Constant Q-Cepstral Coefficients (CQCC) [55]. As these models become larger and are trained on more data, their performance improves [10]. Community-led initiatives, such as the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) challenge [56], independently assess the performance of these classifiers, pushing the state-of-the-art in developing automatic spoof countermeasures. However, a limitation of these approaches is potentially disregarding or insufficiently using human perception in the binary classification. This paper addresses that gap.

The second spoofing countermeasure embeds perceptual or imperceptible watermarks into the initial audio for later authentication either by humans or machines [49]. Recently, watermark usage expanded to identifying AI-generated content across modalities like text, images, and audio [19, 30, 31]. Ideal watermarks withstand manipulations including cropping, compression, or fast-forwarding. Watermarks arise from either hand-crafted or machine-learning approaches. Hand-crafted approaches involve deterministically modifying certain frequency components, like phase or amplitude, to embed identifiable signals without affecting perceptual quality [1]. Machine learning approaches embed imperceptible sound bites using generative models, while separate detection models recognize these embedded signals. For instance, `resemble.ai` leverages psychoacoustics to insert quiet sounds in regions of lower human sensitivity by exploiting proximity with louder sounds in both frequency and time [47]. Our work investigates whether humans can identify implicit audio cues indicative of spoofing. Such perceptual capabilities could inform watermark designs for auditory channels.

3 STUDY 1: UNDERSTANDING HOW BLIND INDIVIDUALS PERCEIVE BONA FIDE & SPOOFED AUDIO

We first conducted an IRB-approved study with 12 blind participants–6 females and 6 males–anonymized as P1 to P12. The study comprised two parts: In the *first* part, we asked participants to classify 63 audio clips as real or fake. These clips were composed of 5 TTS, 8 deep fake, 25 replayed, and 25 bona fide. In the *second* part, we conducted a semi-structured interview to gain insights into their strategies and decision-making processes. This approach enabled us to identify the underlying factors and perceptual cues that blind individuals utilize when tasked with differentiating bona fide and spoofed audio. We curated audio clips from the ASVspoof challenge datasets, a bi-annual competition aimed to advance automatic spoofing countermeasures for audio. Details on participant demographics, dataset curation, study protocol, and data analysis follow next.

3.1 Participants and Recruitment Criteria

We submitted our study protocol and recruitment materials to the Research Advisory Council of the National Federation of the Blind (NFB) in Baltimore, MD, USA. Upon review and approval of our study, they disseminated our recruitment materials (a screen reader-accessible Google Form) to a group of blind participants.

Our recruitment criteria included that participants must 1) be native English speakers; 2) possess headphones; 3) use screen readers, and be under the age of 50, as human hearing sensitivity begins to decline from this age [24]. Our optional recruitment criteria sought participants with a musical background to offer more professional perspectives when listening to audio clips and those who listen to TTS audio at a fast pace, indicating extensive screen reader usage.

Table 1 summarizes participant demographics and headset types. All participants were blind – 7 were blind since birth and 2 had light perception. They represented diverse professional backgrounds, with 3 having musical experience. The listening rates of most participants were over 100% faster than the default playback rate of their screen readers, with 3 at 100% speed. P12 was a notable exception preferring a 60% slower rate. Although she did not explain this

#	Gender	Age	Vision	Onset	Occupation	Screen Reader	Listening Rate	Music Training	Light Perception	Headphone Used
P1	M	20-25	Blind	Since birth	Therapist	VoiceOver	100%	N	Y	EC, WL
P2	M	40-45	Blind	Since 3	Educator	NVDA	130%	N	N	EC, W
P3	M	40-45	Blind	Since 6	IT	JAWS	170%	N	N	RG, W
P4	F	26-30	Blind	Since birth	Student	VoiceOver	170%	N	N	EC, WL
P5	F	40-45	Blind	Since birth	Govt. employee	JAWS	100%	N	N	EC, W
P6	M	26-30	Blind	Since 9	Service	VoiceOver	125%	N	N	EC, WL
P7	F	30-35	Blind	Since birth	IT	VoiceOver	200%	N	N	RG, W
P8	M	30-35	Blind	Since 9	IT	JAWS	170%	Y	N	EC, W
P9	F	30-35	Blind	Since 2010	Govt. employee	JAWS	150%	N	N	EC, W
P10	M	30-35	Blind	Since birth	Service	JAWS	200%	Y	N	EC, W
P11	F	20-25	Blind	Since birth	IT	JAWS	100%	Y	Y	EC, W
P12	F	30-35	Blind	Since birth	Student	JAWS	60%	N	N	EC, W

Table 1: Participants' demographics. Here, EC and RG stand for echo cancellation and regular headphone type, respectively; WL and W stand for wireless (e.g., Bluetooth-enabled) and wired headphone connection, respectively.

Settings \ Settings Code	a	b	c
Acoustic Settings for bonafide clips			
Room Size (sq ft)	2-5	5-10	10-20
Reverberation Time (ms)	50-200	200-600	600-1000
Distance from the microphone (cm)	10-50	50-100	100-150
Recording Settings for spoofed clips			
	A	**B**	**C**
Distance from the microphone (cm)	10-50	50-100	100-150
Audio quality	perfect	high	low

Table 2: Acoustic environments for bona fide audio and recording variables for spoofed audio in the PA dataset. All bona fide clips carry a three-letter lowercase code, like `aaa`, which denotes the settings used for the recording: a room size of 2-5 sq ft (`a`), a reverberation time of 50-200 ms (`a`), and a distance of 10-50 cm from the microphone (`a`). Spoofed clips, on the other hand, feature a five-letter code (e.g., `aaaAA`), beginning with three lowercase letters followed by two uppercase ones that represent the original and manipulated acoustic settings. For example, a spoofed clip coded as `aaaAA` suggests that it originated as a bona fide clip with the code `aaa` and was subsequently replayed and re-recorded at a distance of 10-50 cm (`A`) with 'perfect' audio quality (`A`).

slower rate, we noticed she spoke more slowly than other participants. Except for 2 participants, all used professional-grade echo-cancellation headphones. Two-thirds of these headphones used a wired connection while the remaining one-third used Bluetooth-enabled wireless connection.

3.2 Study Stimuli: Curation of "Challenging" Audio Clips

For this study, we curated "challenging" audio clips from the Automatic Speaker Verification (ASVspoof) Challenge [56]. The challenge organizers provide datasets containing thousands of bona fide and spoofed audio clips, evaluation metrics, and baseline models for comparing competing spoof detection model performance [37, 59].

The curation objectives. Our curation process had two key objectives. First, we aimed to conduct the study one-on-one within 90 minutes. A longer study duration could introduce fatigue as a potential performance-confounding variable given the sustained perceptual focus required. This time limit allowed us to select 60-65 clips under 10 seconds in length, appropriately scoped for in-depth analysis without overtaxing participants.

Second, since the competition datasets contain a large number of audio clips (e.g., over 134,730) spanning numerous bona fide and spoofed configurations (see Table 2), we required a sampling policy to select sufficiently representative yet study-appropriate subsets. This filtering policy enabled meaningful comparisons between human performance and binary classifiers.

The curation policy. Instead of sampling audio clips randomly from the datasets, we first ran a baseline binary classifier, AASERT [36], provided by ASVspoof Challenge, to all audio clips in the datasets and chose those it misclassified. For instance, if the model incorrectly identified a bona fide clip as spoofed, or vice versa, we included it. These selected clips were then sorted by their predicted probability scores, from highest to lowest. A high score for a misclassified clip indicates that the model is highly confident in its incorrect prediction, making the clip "challenging" to classify accurately for the model. We then separated these sorted clips into two categories: *1) False Positives*, where clips were mistakenly predicted as bona fide but were actually spoofed, and *2) False Negatives*, where clips were incorrectly labeled as spoofed but were bona fide.

The curated stimuli. The ASVspoof competition provides three datasets: *Physical Access (**PA**)*, *Logical Access (**LA**)*, and *Deep Fake*

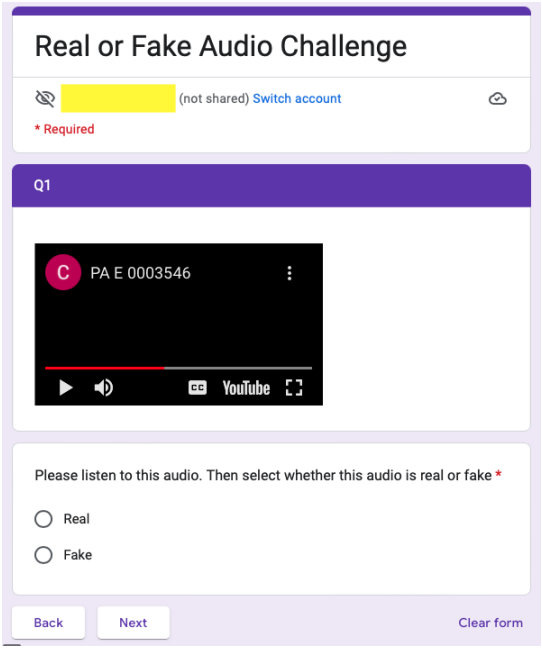

Figure 1: A screenshot of the Google Form used during our study. This form is fully accessible with screen readers and operable with keyboard shortcuts. The URL of this Form is provided in the footnote.

(DF). The PA dataset features both bona fide and spoofed clips. Bona fide clips reflect recordings under varied room acoustics of size, reverberation time, and microphone distance, detailed in Table 2. In contrast, spoofed clips originate from replaying and re-recording bona fide samples under differing recording distances and audio quality, also defined in Table 2. Unlike PA, the LA dataset solely comprises spoofed clips generated via text-to-speech and voice conversion algorithms. Similarly, the 2021 DF dataset presents generative AI-powered spoofed utterances.

We applied our curation policy to all three datasets. In addition to the model's difficulty in making accurate predictions, we also considered other factors such as clips' length, the speakers' gender, speech content variety, and configuration codes. Generally, we opted for longer clips, most falling within the 5-to-7-second range, that covered a range of topics like health, politics, and daily life. The selection also maintained a gender balance, featuring equal numbers of male and female voices. Furthermore, we included at least one clip from each distinct replay configuration, as detailed in Table 2.

Based on the above factors and our curation objectives, we selected the top 25 false positive and top 25 false negative audio clips (50 total) from the PA dataset. Given lower predicted probability scores compared to the PA selections, we selected fewer audio clips from the DF and LA datasets: the top 8 false positives from the DF dataset and the top 5 false positives from the LA dataset. These 63 stimuli fulfill our earlier stated curation objectives.

3.3 Study Apparatus

Next, we embedded these 63 clips in an online Google Form as questions, as depicted in Figure 1, presenting one question at a time to avoid confusion. Each question included two mutually exclusive options—Real or Fake—to indicate the clip's authenticity. Due to Google Forms' inability to directly support audio, we embedded the clips into blank video soundtracks. We ensured the Form's accessibility by testing it with multiple screen readers with different web browsers on Windows and Mac computers. Readers can access the Form from this URL[2].

For counterbalancing, we enabled "shuffle question order" to randomly assign clip order per participant. Furthermore, our use of sub-15-second stimuli, along with auditory sensory memory decay, intrinsically limited inter-stimulus interference as participants progressed through clips.

3.4 Study Procedure and Data Analysis

We conducted the study online using Zoom teleconferencing software. After a brief introduction and obtaining consent, we collected participant demographics, including their history of vision impairment, listening rate and listening devices, education, and employment. Each session was divided into main two parts: an online quiz, followed by a semi-structured interview.

Part 1: an online quiz. We began Part 1 by distributing the Google Form (shown in Figure 1) through Zoom chat or via email based on the participants' preferences. Participants used their personal computers and preferred screen readers while wearing headphones. We advised using Chrome. Once settled, participants shared their screens. They used keyboard commands, such as `Tab`, `Enter`, `Space`, and `Arrow` keys, to play the video clip embedded in the Form, whose audio track contained the audio used in the study. Using these shortcuts, they selected an answer (e.g., `Real` or `Fake`) and navigated to subsequent questions one by one. Participants could replay each clip; we recorded the number of replays as a measure of difficulty. A 60-minute time limit was set to complete this quiz.

Part 2: the semi-structured interview. After form submission, we began Part 2 with a semi-structured interview to understand participants' decision-making in assessing audio clip authenticity. We asked what aspects of the audio they focused on, as well as what factors they prioritized, and what influenced their decisions. Moreover, we asked if there were competing factors, and if so, how they prioritized or broke the tie.

Further, we reviewed their answers one by one without revealing the correct ones and asked them to articulate their decision-making processes for a clip. Additionally, we asked how confident they were in making the decision. For clips that they either spent more time on or replayed more frequently, we probed what were the challenges and their resolution strategies.

We also asked participants to estimate the acoustic settings of the audio, such as the distance from the sound source to the recording device (e.g., close or far), room size (e.g., small, medium, or large), duration of reverberation (e.g., short, medium, or long), and audio quality (e.g., good, medium, bad).

[2]https://forms.gle/YfRtcta6zwbG1aQ47

The first author of this paper conducted the study. Each session lasted approximately 90 minutes. All sessions were recorded post-consent for subsequent analysis. Participants received a $25 Amazon gift card for their participation.

Data analysis. We manually transcribed and analyzed the data using an iterative coding process [13]. All authors reviewed the codebook (created in Google Sheets) during weekly research meetings, identifying key concepts, categorizing them into themes and sub-themes, and resolving any conflicts.

In the first coding cycle, we identified low-level descriptive codes capturing dimensions like speakers' prosody, vocal patterns, accents, mouth sounds, pacing, emotional state, audio quality, distortion, artifacts, ambiance, and post-processing effects. In subsequent cycles, conceptually related codes were consolidated into higher-level categories reflecting core themes. Ultimately, six major themes emerged through the iterative process of refinement and aggregation, detailed in the following sections.

For quantitative analysis, we exported the results from the Google Form to a CSV file and post-processed it in Python. We programmatically compared the participants' Real/Fake classifications to the ground truth labels derived from our curated stimulus datasets.

4 STUDY 1: FINDINGS

We now present our findings from Study 1, organized under six major themes. Results on quantitative prediction accuracy are presented in Table 3 and described toward the end of this section.

4.1 Speakers' Accent

The term accent has two distinct concepts: i) voice inflection, which refers to the specific emphasis a speaker applies to a syllable or word in speech through stress or pitch; ii) it relates to the distinctive manner in which a speaker pronounces a language, often linked to a particular nation, locality, or social class.

4.1.1 Voice Inflection. All participants attended to the speakers' vocal inflection, intonation, fluctuations, variations in sound, and vocal tone in the audio clips. This attribute emerged as the most commonly mentioned (N=12), with participants focusing on the extent of vocal fluctuations and monotonicity to discern whether the voice belonged to a real human.

When encountering keywords in a sentence during the speech, participants directed their attention toward emphasis, specifically how the speakers emphasized particular words, made their voice more assertive, or elevated their pitch. Bona fide human speakers assign importance to certain syllables within a sentence and individual words, which they accomplish through inflection. This inherent quality guided our participants to select "real" in their responses.

Moreover, participants noted that changes in tone at the end of a sentence served as a significant indicator. Screen readers like VoiceOver or JAWS attempt to deliver the context of a sentence clearly, resulting in less variation in vocal tone compared to actual human speakers. Consequently, participants perceived relatively monotonous voices as spoofed audio. P11 stated that machines only elevate the tone at a sentence's end when it terminates with an exclamation (!) or a question mark (?). In contrast, humans may occasionally interpret the sentence as a question, even in the absence of a question mark or when comprehension is incomplete.

The speaker's overall pitch also influenced participants' choices. P9 cited an instance where the voice did not align with their preexisting concept of a typical male or female voice. Upon hearing an exceptionally low or high-pitched voice, P9 assumed it belonged to a natural person with an atypical pitch, as screen readers like VoiceOver or JAWS do not offer voices with such distinctive pitches. However, this observation does not necessarily imply that a voice adhering to the prototypical human voice they envision is spoofed. Furthermore, speakers' gender did not influence participants' decisions (N=7).

4.1.2 International Accent. Regarding the second definition of accent, the majority of participants preferred to set their screen reader to have a North East American accent voice (N=11). Accordingly, they are familiar with listening to accents of the north-east states of the USA but unfamiliar with other accents, such as the accent of the southern states of the USA or international (e.g., European and Asian) accents. Although companies offering screen reader services attempt to design voices with non-American accents, these options are infrequently provided or used due to limited demand. Therefore, when participants encountered international accents, such as European or Asian accents, they tended to perceive those audio clips as bona fide. This is perhaps due to their misconception that machines cannot learn different accents easily.

> *"'S' sounds different. When she speaks 'lists', she sounds like a European, you know how Spanish speaks 'S'. That's why I chose real."* (P2)

> *"It sounds like a foreigner. I mean, non-American accent. There is a synthesized voice with a non-American accent but you can usually tell the difference [between synthesized voice and real human voice] with how the inflection is or how it handles the accents."* (P8)

Additionally, participants interpreted unusual pronunciation as errors made by humans. For instance, P1 posited that people read acronyms alphabetically or made mispronunciations according to their discretion, in contrast to the pronunciation provided by screen readers. This finding is fascinating—it reiterates the age-old saying, *"to err is human"*.

In sum, participants effectively differentiated between real human voices and synthesized voices (spoofed by default) by listening to international accents, unfamiliar accents, and international pronunciations.

4.2 Human Sound

Participants identified bona fide and spoofed audio clips by the presence or absence of unique human sounds. These include breathing, mouth movements like lip-smacking, filler words such as *hmm* and *ah*, and vocal fry sounds.

4.2.1 Breathing sound. For the majority of participants (N=10), the presence of breathing sounds at the beginning of the audio clip was a vital factor. If no breathing sounds were present, they considered the audio to be synthetic. In the case of replayed audio clips, breathing sounds may be difficult to detect since they are

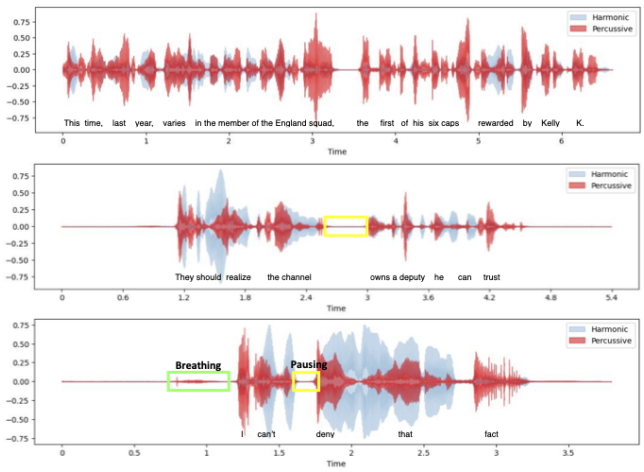

Figure 2: Top: Waveplot of a TTS clip. Middle: Waveplot of a spoofed clip with a replay attack. Bottom: Waveplot of a bona fide clip which is the original version of a human voice.

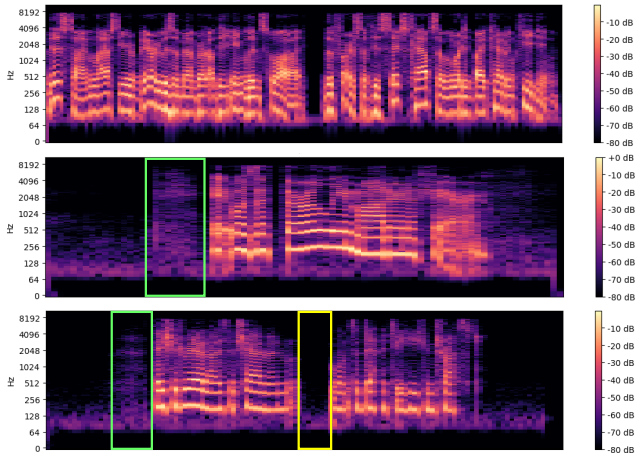

Figure 3: Top: Spectrogram of a TTS clip. Middle: Spectrogram of a spoofed clip with a replay attack. Bottom: Spectrogram of a bona fide clip which is the original version of a human voice.

often too faint to be heard during replay. However, in original clips, the presence of breathing sounds at the beginning is more likely, which helped participants correctly identify these clips as bona fide.

To visually demonstrate this factor, we display wave plots (Figure 2) and spectrograms (Figure 3) of three audio clips. In wave plots, red-colored waves are percussive, which represent high-pitch and non-tonal noises. These are used to replicate the articulation of natural speech. Blue-colored waves are harmonics that illustrate the intricate interaction of the vocal cords, throat, or nasal cavity that forms the distinctive timbre of the voice.

The middle plot, representing a replayed clip, shows no significant voice activity at the start. In contrast, the bottom plot (Figure 2), displaying a bona fide clip, exhibits distinguishable energy movement, highlighted in the green box.

4.2.2 Filler words. P7 highlighted that machines typically interpret sentences inputted into a computer differently from the informal language individuals utilize in speech. She further noted that filler words, such as *hmm...* and *ah...*, are rarely incorporated in written sentences, yet they are frequently employed by genuine human speakers during conversations. Hence, when participants detect filler words in audio clips, they can deduce that the voices belong to real humans.

> *"I heard 'Rrrr'. Also, vocal registers, that make the speaker sound like very intelligent. Those matter. Humans do that while screen readers don't."* (P7)

Filler words usually have more air than regular words and are frequently used during hesitation, resulting in instances where the volume is low within the clip. Similar to breathing sounds and mouth movements, participants mentioned that filler words might become faint when clips are replayed multiple times and re-recorded. This provided them with another clue to correctly identify these clips as spoofed (e.g., clips that contain faint filler words are likely to be spoofed).

4.3 Pausing

Participants took into account the location and duration of pauses when determining whether an audio clip was bona fide or spoofed. They noticed many artificial or nonsensical pauses in synthesized voices and overlapping of words (N=11). However, P7 noted that sometimes overlapping of words can still resemble a human voice if the length of the pause is typical. P10 also mentioned cases where a person's breathing sound overlapped with a word without an appropriate pause, which they perceived as an added breathing effect to imitate a human voice. Overall, participants tended to consider an audio clip as spoofed if they noticed an abnormally short pause.

> *"If you breathe, you at least talk after you breathe, but if there is no space after a breath or a raspy kind of breathing, that is fake. Like if someone is pretending to be another, or a person might constantly start to talk after breathing. But I don't know anyone who breathes like that and who can talk at the same time. People breathe until they can talk."* (P10)

Moreover, P4 explained that she could recognize choppiness and strange silence which was not expected unless the pronunciation was mistaken. She considered it as a spoofed audio created by concatenating two different files to form a new sentence. For example, the word 'overlay' would typically be pronounced as a single unit, but one of the given clips enunciated it as if 'over' and 'lay' were two separate words. This led to cases where the entire clip was perceived as spoofed.

In the context of an authentic human voice, the participants mentioned that the duration of pauses between words within a sentence is inconsistent. For example, the pause following specific punctuation marks, such as commas, semicolons, or colons, generally exceeds that of the pause after a word. Furthermore, humans, unlike text-to-speech machines, regulate speech speed and pause

in accordance with a unit of meaning. While machines interpret spaces literally, humans read or construct spaces based on the unit of meaning within a sentence to facilitate comprehension. Therefore, the pattern of human pauses appears irregular in comparison to synthesized voices, enabling listeners to discern a natural flow or rhythm in the audio when reading scripts.

The noted differences are discernible in our previous visualization, Figure 2 and Figure 3. The most prominent distinction between TTS and replayed audio lies in the presence or absence of hesitation at sentence commencement. TTS starts without delay, evidenced by the absence of gaps between the clip's start and the first word in both figures. Conversely, bona fide and replayed human voices exhibit silence in the beginning.

4.4 Perceived Emotion

The majority of participants (N=8) mentioned that they were able to perceive emotions from the audio clips. Although the types of emotions varied, they commonly agreed that various emotions, such as sadness, matter-of-fact tones, boredom, and excitement, were evident in actual human voices. Furthermore, participants identified similar emotions when listening to the same specific audio clips.

> *"I felt her sadness in her tone, some kind of emotion. Usually, I can't feel any emotion from synthesized voice."* (P2)

Participants reported perceiving various emotions, such as happiness and sadness, and attempts to conceal emotions. Many commented that the speaker seemed bored, speculating that the individuals recording the sentences might have been fatigued by the repetitive nature of the task. However, participants were still able to discern various emotions beyond boredom that machines could not replicate, leading them to correctly classify these clips as real.

However, when listening to spoofed audio clips, participants reported only three emotions—*no emotion*, *boredom*, and *anger*—all of which were negative (N=6). Specifically, they struggled to identify the type of emotion conveyed in spoofed audio clips. P9 mentioned that replayed (spoofed) audio clips sounded as though there were numerous layers between the recorder and the speaker, making it challenging to detect any emotions. TTS voices did convey some emotions, but they were primarily negative. Three participants reported feeling anger from machine-generated voices due to the lack of natural reverberation and monotonous tone. In all instances where emotions influenced decision-making, participants made accurate judgments.

> *"The one that lady was talking about the meeting [Real], I was able to feel emotion from it. I felt sad from the way she is speaking and the content as well. But definitely from the way she is speaking. That is how I try to tell people's emotions by the tone of their voice. The person who was talking about the meeting [Real] had a tighter voice to her voice. From the synthesized voice [Fake], no emotion at all."* (P12)

The audio content also influenced the perception of emotions. When P12 heard clips discussing societal issues with accurate pronunciation, she associated the voice with a news report. If the topic was relevant to the news, it gave her the impression that the voice was confident and trustworthy. However, in this case, the feeling was not attributed to the audio aspect but rather the content. This suggests that content can also be associated with emotions. Thus, content analysis could be combined with emotion detection models to determine whether a voice is genuinely human or an impersonation.

4.5 Audio Quality, Digital Artifacts, and Echo

Some distinct sounds are exclusively present in spoofed audio recordings. Participants described these as "mechanical sounds", "digitized sounds", "metallic sounds", "grainy sounds", and "fish tank sounds". They considered these sounds to be defining characteristics of screen reader voices or digital artifacts introduced during spoofing. Consequently, when voices exhibited these types of artifacts, participants correctly identified them as spoofed. In Figure 2, red-colored percussive waves, which indicate clicks, pops, and other types of noise, are more dominant in TTS and replayed clips compared to the bona fide ones.

> *"It is a screen reader. I think I've heard this voice a lot. It sounds really robotic and familiar."* (P6)

> *"It is obviously a synthesized voice. I can't hear anything that sounds like a human. It has really robotic, mechanical voice. It has no emotion at all. No breathing, no echo, no anything, not even white background noise. It is just like text-to-speech, which I used to listen."* (P11)

Regarding sound quality, participants (N=9) observed that the presence of echo influenced their decision-making. The absence of echo led them to perceive the audio as synthesized, as they believed this to be a common characteristic of synthesized voices. Even if the voice itself sounded human, the lack of echo led participants to think that the audio had been manipulated to remove background noise.

Additionally, highly static audio with strong reverberation was also perceived as manipulated, but in this case, to add background noise. On the other hand, excessive echo led participants to believe that the audio had been replayed and re-recorded multiple times, causing the quality to deteriorate. In contrast, if there was an appropriate level of reverberation, they regarded it as natural white noise and considered the audio to be bona fide. According to P4:

> *"Replayed audio clips sound like they are in the fish tank or the sound comes from the drive-through kiosk machine. It means these sounds give me the impression that there are several layers on the voice, but this is different from the original clips that are recorded from a long distance. Quality is so bad."*

Similar to emotion, participants perceive sound quality in relation to sentence content. They tend to perceive a voice as neat and as originating from a real conversation when the sentence deals with everyday topics or familiar subjects, and they can easily understand it due to its high-quality sound. However, if the topic is too complex or the sentence sounds disorganized, participants may believe the sound quality is low or that the sentence was recorded

by a machine rather than spoken by a real person. Thus, it is challenging to prevent personal interests, biases, or stereotypes about everyday conversations from influencing individuals' decisions.

4.6 Analysis of Predicted Accuracy

The overall prediction accuracy for blind participants was 57% (see Table 3). This low accuracy suggests that, in general, blind participants struggled to correctly determine the authenticity of an audio clip. Upon itemized analysis, we found that blind participants were highly accurate (accuracy: over 90%) in detecting text-to-speech (TTS) audio. Their greater use of screen readers and TTS audio may have contributed to this performance. Apart from TTS, blind participants were most accurate in detecting bona fide clips (accuracy: 64%), much higher than recognizing deep fake audio (accuracy: 58%) or replayed (accuracy: 43%). Finally, we did not observe any effect of headphone types (e.g., echo-cancellation or regular) or headphone connections (e.g., wireless or wired).

We emphasize that the overall accuracy of 57% was still significantly higher than the deep-neural network-based baseline binary classifier, ASSERT [36]. Recall that the audio clips used in the study were "challenging" for the baseline classifier (Section 3.2), which misclassified them all with high confidence, yielding an accuracy of 0%. Therefore, we conclude that human traits in the audio, as well as the context of the audio, are instrumental for better prediction. Put differently, machine learning models can still benefit from human judgments in classifying challenging instances. We elaborate more on this possibility in the Discussion section.

4.6.1 TTS as the Ground Truth for Spoofed Audio. A deeper analysis reveals a key insight: throughout the study, blind participants compared an audio clip they had just listened to with their familiar TTS speech, mentally extracting the human traits and expressed emotions in the audio (if any), assuming TTS as the ground truth for spoofed audio that is devoid of any human traits or emotion. This explains their high accuracy (over 90%) in detecting TTS.

A side effect of this assumption was that if the audio contained a natural human voice during its creation, as in the audio clips in PA datasets, but was then spoofed by replaying or re-recording it in various acoustic settings, participants were biased towards accepting it as bona fide. They placed less emphasis on the presence of recording artifacts in the audio. This accounts for the lowest accuracy (43%) in detecting replayed clips, making it the most challenging spoofing type for blind individuals to detect.

4.6.2 High-Frequency Vocal Sound in Deep Fake Speech. Three musically trained participants (P8, P10, and P11) noticed the presence of certain high-frequency speech elements, which differed from TTS or natural human speech. Surprisingly, their observation was consistent with prior work on neural-rendered audio — the presence of higher frequencies, particularly in vocals, is a common limitation in many deep fake audio synthesis models (e.g., MelGAN [34]; see Frank et al. [21] for additional details). Other participants also detected something unusual about the deep fake audio but were less articulate. Eventually, many were swayed by the occasional presence of human-like traits, such as vocal inflection, intonation, and fluctuations. As deep fake algorithms continue to evolve, our findings suggest that blind participants are likely to categorize them as bona fide. This underscores the importance of audio watermarking for AI-generated audio. We delve further into this issue in the Discussion section.

5 STUDY 2: A SURVEY WITH SIGHTED AND BLIND INDIVIDUALS

Building on the insights from Study 1 regarding the decision-making process of blind individuals, we aimed to determine whether sighted individuals employ the same process and, if not, how their decision-making diverges (RQ2).

To investigate this, we conducted an IRB-approved online survey. We more than doubled the number of blind participants (from 12 to 30) and included an equal number of sighted participants in Study 2. Additionally, we increased the total number of clips from 63 to 96 to enhance statistical power. The clips consisted of 8 TTS audio clips, 8 deep fake (DF) audio clips, 40 replayed audio clips, and 40 bona fide clips. Following a method similar to Study 1, detailed in Section 3.2, we integrated these clips into a quiz-like Google Form, similar to Figure 1. Among nine potential combinations of replayed audio, we included four in the study because those samples were challenging to distinguish, having higher false positive predicted probability scores from the binary classifier.

At the end of our Google Form, we incorporated a textbox where participants could describe their decision-making process. The label prompted participants to reflect on differing configurations and origins of an audio clip, providing examples like the distance between the microphone and the sound source, room sizes, replay settings, synthesis methods (TTS and deep fake), and how these variations influenced their decisions.

We distributed the URL of this Form to online forums for blind and low-vision users (e.g., nvda@nvda.groups.io and program-l@freelists.org). To enlist sighted participants, we shared the URL via university mailing lists. Participation was anonymous and voluntary, and we offered no financial incentives. Unlike in Study 1, we had no direct interaction with the participants in Study 2. Beyond individuals' responses to the 96 audio clips and the free-form text response at the end of the form, we collected additional open-ended feedback through emails from some participants.

Data Analysis. We analyzed the responses to audio clips based on their accuracy scores (high true positive, high false positive, high false negative) and the predominant group with correct answers (sighted, blind, or an equal split). Additionally, we explored common features in audio snippets that were either easily identifiable or challenging for both sighted and blind individuals.

All participants provided free-text responses of varying lengths. For text analysis, we utilized the existing theme codebook developed in Study 1. Through iterative passes, we tagged themes already captured in the codebook while noting newly emergent ones as well. One such new theme highlighted how sighted respondents envisioned speakers' facial movements while clips played. We now present our findings in the next section.

Audio Category	Study 1		Study 2		
	Participants: 12 Blind		Participants:	30 Blind	30 Sighted
	Number of clips	**Mean Accuracy (SD)**	**Number of clips**	**Mean Accuracy (SD)**	**Mean Accuracy (SD)**
Bona Fide	*25*	*0.64 (0.16)*	*40*	***0.67 (0.16)***	*0.65 (0.15)*
TTS	5	0.97 (0.06)	8	**0.91 (0.05)***	0.84 (0.08)
Deep Fake	8	0.58 (0.12)	8	0.58 (0.12)	**0.71 (0.17)***
Replayed	25	0.43 (0.17)	40	0.44 (0.18)	**0.46 (0.15)**
Overall	**63**	**0.57 (0.22)**	**96**	**0.59 (0.21)**	**0.59 (0.29)**

Table 3: The average accuracy and standard deviation (SD) for identifying bona fide and spoofed audio types-—including TTS, Deep Fake, and Replayed -— in Study 1 (Section 3) and Study 2 (Section 5). In Study 2, numbers marked with an asterisk (*) indicate statistically significant differences in accuracy between blind and sighted participants (by Mann-Whitney U tests).

5.1 Findings: Comparing Decision Factors of Sighted and Blind Individuals (RQ2)

Our survey results, summarized on the right side of Table 3, confirm the trends identified in Study 1. Sighted individuals also attended to similar human traits in audio, underscoring the validity and significance of our earlier findings. However, we observed a crucial difference in the mental models that sighted and blind participants employ for decision-making, detailed below.

5.1.1 Differences in Mental Models between Sighted and Blind Participants. Study 1 revealed blind participants typically compare incoming audio to text-to-speech (TTS), using TTS as a spoofed audio baseline. For sighted participants, our text analysis revealed that they envision an abstract human face, like a talking head, as well as its facial expressions that would likely accompany speech production. One sighted participant wrote the following:

> *"While listening, I can imagine a girl of a specific age, and I begin to draw a sense of connection to someone in a similar age group and gender."*

Another sighted participant wrote: *"How could she make that sound? She must have dropped her jaw and twisted her tongue."*

One way to explain this phenomenon is that sighted individuals engage with their communication partners visually, often paying close attention to each other's facial expressions and mouth movements. Thus, when listening to a speech clip, they try to visualize the likely mouth movements and facial expressions involved in the speech. If these imagined mouth movements and/or facial expressions seem unrealistic, they tend to consider the audio as deep-faked.

In sum, blind participants mentally assess whether the audio exhibits human traits that are absent in TTS, while sighted participants evaluate whether the speech could realistically be produced by an abstract talking head. This variance in mental models is key to explaining the performance differences between sighted and blind participants.

5.1.2 Analysis of Predicted Accuracy. For each audio clip, we collected 60 responses—30 from blind individuals and 30 from sighted ones—resulting in a total of 5,760 data points (60×96). To assess statistical significance, we used the Mann-Whitney U test, a non-parametric method, given that our data were not normally distributed.

Overall, the mean accuracy scores for blind and sighted participants were almost identical – 59% (SD: 21%) and 59% (SD: 29%), respectively, as shown in Table 3. The mean accuracy scores differed only in the third decimal place.

Our itemized analysis indicated that sighted participants were more accurate (71%) in detecting deep fake audio than their blind counterparts (58%), a statistically significant difference. Conversely, blind participants significantly outperformed sighted ones in recognizing TTS, with scores of 91% versus 84%. These performance gaps can be attributed to the differing mental models outlined earlier.

In recognizing bona fide audio, blind participants performed marginally better than sighted participants (67% vs. 65%), although the difference was not statistically significant. The performance of both groups was the weakest in recognizing replayed audio, with scores of 44% and 46%, making it the most challenging category of spoofed audio to detect. Note that this is consistent with our findings in Study 1.

5.1.3 Analysis of Replayed Audio. Since replayed audio appeared as the most challenging category, we performed an itemized analysis of these clips. However, we only report the dominant trends in the data, as our data points are insufficient compared to the large number ($243 = 27 \times 9$) of possible replay configurations.

We found the distance from the microphone during recordings was a key contributor to this challenge. Blind individuals exhibited higher accuracy when the audio was replayed from a longer distance and the perceived sound distance was distant. When audio was replayed at a closer range, implying that the sound was perceived to be near, sighted individuals demonstrated greater accuracy. This is surprisingly consistent with prior work that reported that blind individuals' judgment improves with sound from longer distances, while sighted individuals discern sounds more accurately when they originate from closer proximity [33]. Finally, regarding re-recording in smaller room sizes, both blind and sighted individuals appeared to be perplexed, recording the lowest accuracy scores for both groups.

Human Trait	Description	Likely Decision
Accent (Inflection)	Obvious inflection throughout the clip	Bona fide
Accent (International)	European or Asian English accent	Bona fide
Audio quality	Too Good: All words are clearly understandable, with no ambiance noise, background sound, reverberation, or echo	Spoofed
	Medium: Noisy but understandable	Bona fide
	Bad: Too much noise and words are not understandable	Spoofed
	Metallic sound present anywhere in the audio	Spoofed
Breathing	Breathing present at the beginning	Bona fide
	Breathing present at the end	Bona fide
	Breathing anywhere in the audio	Bona fide
Echo	Echo present at the end	Spoofed
Emotion (Negative)	Evoked negative emotion	Spoofed
Emotion (Neutral)	Evoked no emotion	Spoofed
Emotion (Positive)	Evoked positive emotion	Bona fide
Filler words	Noticeable filler words present at least once	Bona fide
Human Error	Mispronunciation present at least once	Bona fide
Pausing	Pausing present at the beginning	Bona fide
	Pausing anywhere in the audio	Bona fide
Reverberation	Reverberation present anywhere in the audio	Bona fide
	Reverberation present at the end	Spoofed

Table 4: A comprehensive list of the human traits that may have affected the choices of blind participants in our study.

6 DISCUSSION AND DESIGN IMPLICATIONS

6.1 Summary of Findings and Implications

Blind individuals attend to specific human traits in the audio, which we summarized in Table 4, along with how these could influence their decisions. Further, blind individuals were highly accurate in detecting TTS audio, with accuracy exceeding 90%. Moreover, TTS serves as their mental model for evaluating audio authenticity. Finally, the inclusion of human-like traits in spoofed audio types, such as replayed or deep fake, complicates their ability to correctly identify authentic audio. Additionally, our second study doubled the participant pool including both blind and sighted individuals, collecting more data points. It further generalized or confirmed trends observed in the first study for both blind and sighted individuals. For example, sighted participants, like their blind counterparts, focus on human traits in audio to determine authenticity. However, the mental model of sighted individuals differs from blind individuals – sighted individuals rely more on visualized facial expressions during speech as their decision-making baseline. Because of this difference, blind participants outperformed sighted ones in detecting TTS audio (91% vs. 84%) but were less adept at recognizing deep fake audio (58% vs. 71%). Finally, both groups struggled to accurately identify replayed audio. When the audio was replayed at a closer range, the sighted group demonstrated greater accuracy. In contrast, blind individuals exhibited higher accuracy when the audio was replayed from a longer distance. The findings have major implications which we discuss next.

6.2 Personal and Societal Implications

First, when an audio clip features human traits, even if it is generated by an AI, blind individuals are likely to perceive it as authentic. This tendency puts blind individuals at risk across voice interfaces like telebanking and bill pay as convincing AI personas proliferate.

Second, the judicious application of realistic deep fake voices in screen readers presents promise – substituting the traditional robotic TTS voices with human-like ones can improve the user experience. Particularly, it can make the interaction more affective, personalized, and enjoyable than the status quo. A recent study [51] supports this hypothesis – people like a voice assistant that sounds more like them than one with a less similar voice.

Third, the audiovisual correspondence is strong among sighted individuals [38], providing them with more cues to detect AI-generated audio.

Fourth, the mere fact that audio can be spoofed negatively affects both blind and sighted people's judgment. As they scrutinize the authenticity of the audio more, they become more prone to errors, with accuracy in determining bona fide audio under 65%. This suggests that the task is psychologically taxing and burdensome. This implies that in the future, tasks should not require users to vet the authenticity of the other communication entity based on their speech. Put differently, security measures should be decoupled from audio-based communication; audio can serve other collaborative aims but must not remain tightly coupled to identity verification.

Finally, both blind and sighted individuals apply social norms and judgments in determining the authenticity of the audio, as if conversing with another human in social settings, attending to physiological and emotional states. Since synthetic audio is likely to exhibit more human traits, the binary question of real versus fake becomes less relevant. Instead, research and design should focus on developing better disclosure mechanisms. For example, *is it sufficient to simply disclose that the other party is an AI? Does it affect the communication quality? How to build new paradigms*

embracing both human and synthetic interlocutors transparently? We leave these inquiries for future research.

6.3 Implications in Human Traits Detection

Historically, research on human speech and voice has focused on inferring speakers' personalities [5]. While this line of inquiry determined that specific voices could evoke stereotyped personality judgments, it also revealed that these judgments might not correspond with more direct or valid personality evaluations [20]. Recent studies have shifted their focus from real humans to virtual agents, such as online bots and social spambots, aiming to identify various human traits (e.g., age, gender, personality types, sentiment in generated text, expressed emotions) and specific stereotypes associated with these agents [23]. Their findings suggest that social spambots display limited gender, age, and emotional variation but exhibit higher levels of positive sentiment and usage of neurotic language compared to real humans.

Our work aligns with this direction, as we discovered that both bona fide and spoofed audio could exhibit certain human traits and evoke specific stereotypes among blind individuals. For example, spoofed audio predominantly evokes negative emotions, similar to social spambots that mainly use neurotic language.

While detecting emotions in the text is well-researched, isolating human traits like breathing, pausing, and affect from audio remains less explored. One of our findings is that human judgment still outperforms current binary classifiers for detecting audio authenticity. This implies that segmenting human traits presents an opportunity to construct specialized classifiers, with each focused on identifying a single trait. By combining these individual classifiers into an ensemble architecture, more robust, explainable, and perceptually grounded defenses against spoofing attacks may emerge. Table 4 can serve as guidelines for constituting such classifiers.

6.4 Implications in Human-AI Collaboration

When humans and AI possess identical information, AI tends to outperform human decision-making. However, our findings reveal that in detecting the authenticity of audio, humans possess insights that AI models lack. In such scenarios, a combined approach involving both humans and AI may yield superior performance [15].

One less-explored avenue for enhancing effective human-AI interaction is the use of AI Uncertainty Quantification in predictions. For instance, if a classifier reports its level of uncertainty, such as *"the audio is spoofed with 80% certainty"* or *"the audio is spoofed with 50% certainty"*, it will likely influence the user of this algorithm in their decision-making [4].

Building on this notion of uncertainty, we propose a *deferral learning architecture* [40, 46] – if the model's uncertainty falls below a certain threshold, it defers the decision-making to humans. Prior work has shown that models that learn to defer outperform either humans or AI acting alone [46, 61].

The key challenge to developing this architecture is that it will require a substantial volume of expert judgments to pinpoint instances that should be deferred to human experts. Future employment opportunities for stakeholders who are interested in serving as experts may become available and present an opportunity for blind people who are underemployed [25]. Through the implementation of an online interface, these individuals can contribute efficiently and consistently as crowdworkers, enhancing both employment opportunities for blind people and AI security.

6.5 Implications in Digital Watermarking

Generating an imperceptible watermark is an active area of research, dominated by big technology companies [2, 3, 47]. However, our findings implicate that watermarks need to be perceptible to humans, easy to recognize, and should be a part of the disclosure mechanism. This way, it can promote human agency and control over AI models that encode or decode watermarks imperceptibly.

As such, we encourage researchers to explore human-centered "perceptible" watermarks. A possible idea is to replace a small portion of deep fake audio with a TTS-synthesized speech that blind individuals can instantly recognize. Another idea is to remove breathing sounds temporarily, which both blind and sighted individuals can immediately notice, which will inform them that audio is machine-generated. We leave this work for the future.

6.6 Limitations

Our work has several limitations. First, the imbalanced number of samples across spoofing categories is a limitation, as certain categories (e.g., replayed) were more prevalent than others (e.g., TTS or deepfake) due to our design choice of selecting challenging clips. Second, within the replayed category, some configurations featured only a few samples (e.g., 1 or 3 clips) because of the large number of possible configurations for how an audio clip can be replayed. Finally, the number of samples in the text-to-speech and deepfake categories was small. We addressed these limitations by emphasizing the dominant trends in the results and reporting descriptive statistics. We also stress that our chosen samples were challenging and therefore more informative than arbitrarily chosen samples. In the future, we plan to conduct a large-scale study of deepfake and text-to-speech categories in the wild, as these types of audio clips are becoming more prevalent due to the ubiquity of AI technology.

7 CONCLUSION

We explored the ability of blind and sighted individuals to differentiate bona fide and spoofed audio files and investigated the factors that contribute to their decision-making process. To do so, we first conducted interviews with 12 blind participants, analyzing 63 challenging audio clips. This unveiled specific human traits on which they focus when determining the authenticity of an audio clip. These heuristic traits include the speaker's accents, vocal inflections, the presence of breathing sounds, lip movement sounds, irregular pausing patterns, audio quality, and perceived emotions. Afterward, we conducted a survey with 30 participants from both the blind and sighted groups, respectively. This strengthened the results from the interviews, showing that people rely on human traits for making decisions. However, the ways they used those traits differed: blind people compared traits from screen readers to the given audio samples, while sighted people compared real human voices to the given audio samples. Additionally, audio recording settings such as room size and recording distance impacted the results. Furthermore, our findings provide insights for enhancing AI's spoofing detection capabilities by emphasizing the role of human auditory

understanding in distinguishing between bona fide and spoofed audio. As AI-generated audio becomes increasingly sophisticated, we suggest that the focus should shift from origin determination to risk mitigation, possibly through perceptible watermarks.

ACKNOWLEDGMENTS

We thank sighted and blind participants for their cooperation in this work. We also thank the anonymous reviewers for their insightful comments. This research was partially supported by the US National Institutes of Health and, the National Library of Medicine (R01 LM013330).

REFERENCES

[1] 2023. GitHub - ShieldMnt/invisible-watermark: python library for invisible image watermark (blind image watermark) — github.com. https://github.com/ShieldMnt/invisible-watermark. [Accessed 11-09-2023].

[2] 2023. Seamless Communication Models - AI at Meta — ai.meta.com. https://ai.meta.com/resources/models-and-libraries/seamless-communication-models/#safetyandresponsibility. [Accessed 13-12-2023].

[3] 2023. Transforming the future of music creation — deepmind.google. https://deepmind.google/discover/blog/transforming-the-future-of-music-creation/. [Accessed 13-12-2023].

[4] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion* 76 (2021), 243–297.

[5] David W Addington. 1968. The relationship of selected vocal characteristics to personality perception. (1968).

[6] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. 2020. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 1–6.

[7] Gerry TM Altmann. 1995. *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. Mit Press.

[8] Amir Amedi. 2004. *Visual and multisensory processing and plasticity in the human brain*. Hebrew University of Jerusalem.

[9] Syed Masum Billah. 2019. *Transforming assistive technologies from the ground up for people with vision impairments*. Ph. D. Dissertation. State University of New York at Stony Brook.

[10] Jean-Francois Bonastre, Hector Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Paul-Gauthier Noe, Jose Patino, Md Sahidullah, et al. 2021. Benchmarking and challenges in security and privacy for voice biometrics. *arXiv preprint arXiv:2109.00281* (2021).

[11] Danielle Bragg, Cynthia Bennett, Katharina Reinecke, and Richard Ladner. 2018. A Large Inclusive Study of Human Listening Rates *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174018

[12] Danielle Bragg, Katharina Reinecke, and Richard E Ladner. 2021. Expanding a Large Inclusive Study of Human Listening Rates. *ACM Transactions on Accessible Computing (TACCESS)* 14, 3 (2021), 1–26.

[13] A. Bryman and R.G. Burgess. 1994. *Analyzing Qualitative Data*. Routledge. https://books.google.com/books?id=KQkotSd9YWkC

[14] Ronald A Cole. 2016. *Perception and production of fluent speech*. Routledge.

[15] Mary Missy Cummings. 2014. Man versus machine or man+ machine? *IEEE Intelligent Systems* 29, 5 (2014), 62–69.

[16] Delphine Dahan. 2010. The time course of interpretation in speech comprehension. *Current Directions in Psychological Science* 19, 2 (2010), 121–126.

[17] Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing* 28, 4 (1980), 357–366.

[18] Elevenlabs. 2023. ElevenLabs || Prime Voice AI — beta.elevenlabs.io. https://beta.elevenlabs.io. [Accessed 03-May-2023].

[19] Hany Farid. 2023. Watermarking ChatGPT, DALL-E and other generative AIs could help protect against fraud and misinformation — theconversation.com. https://theconversation.com/watermarking-chatgpt-dall-e-and-other-generative-ais-could-help-protect-against-fraud-and-misinformation-202293. [Accessed 16-08-2023].

[20] Paul J Fay and WARREN C MIDDLETON. 1939. Judgment of Spranger personality types from the voice as transmitted over a public address system. *Journal of Personality* 8, 2 (1939), 144–155.

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

[22] Candice R Gerstner and Hany Farid. 2022. Detecting real-time deep-fake videos using active illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 53–60.

[23] Salvatore Giorgi, Lyle Ungar, and H Andrew Schwartz. 2021. Characterizing Social Spambots by their Human Traits. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 5148–5158.

[24] Sandra Gordon-Salant. 2005. Hearing loss and aging: new research findings and clinical implications. *Journal of Rehabilitation Research & Development* 42 (2005).

[25] Brien A Holden. 2007. Blindness and poverty: a tragic combination. , 401–403 pages.

[26] Jonggi Hong and Leah Findlater. 2018. Identifying speech input errors through audio-only interaction. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[27] Jonggi Hong, Christine Vaing, Hernisa Kacorri, and Leah Findlater. 2020. Reviewing Speech Input with Audio: Differences between Blind and Sighted Users. *ACM Transactions on Accessible Computing (TACCESS)* 13, 1 (2020), 1–28.

[28] Kirsten Hötting and Brigitte Röder. 2009. Auditory and auditory-tactile processing in congenitally blind humans. 258, 1-2 (dec 2009), 165–174. https://doi.org/10.1016/j.heares.2009.07.012

[29] Anil K Jain, Arun Ross, and Sharath Pankanti. 2006. Biometrics: a tool for information security. *IEEE transactions on information forensics and security* 1, 2 (2006), 125–143.

[30] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. 2023. Evading Watermark based Detection of AI-Generated Content. *arXiv preprint arXiv:2305.03807* (2023).

[31] Makena Kelly. 2023. Meta, Google, and OpenAI promise the White House they'll develop AI responsibly — theverge.com. https://www.theverge.com/2023/7/21/23802274/artificial-intelligence-meta-google-openai-white-house-security-safety. [Accessed 11-09-2023].

[32] Tomi Kinnunen and Haizhou Li. 2010. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication* 52, 1 (2010), 12–40.

[33] Andrew J Kolarik, Rajiv Raman, Brian CJ Moore, Silvia Cirstea, Sarika Gopalakrishnan, and Shahina Pardhan. 2020. The accuracy of auditory spatial judgments in the visually impaired is dependent on sound source distance. *Scientific reports* 10, 1 (2020), 7169.

[34] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems* 32 (2019).

[35] Mari Ganesh Kumar, Suvidha Rupesh Kumar, M. S. Saranya, B. Bharathi, and Hema A. Murthy. 2019. Spoof Detection Using Time-Delay Shallow Neural Network and Feature Switching. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2019), 1011–1017.

[36] Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak. 2019. ASSERT: Anti-spoofing with squeeze-excitation and residual networks. *arXiv preprint arXiv:1904.01120* (2019).

[37] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al. 2022. ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild. *arXiv preprint arXiv:2210.02437* (2022).

[38] John MacDonald and Harry McGurk. 1978. Visual influences on speech perception processes. *Perception & psychophysics* 24, 3 (1978), 253–257.

[39] William D Marslen-Wilson. 1987. Functional parallelism in spoken word-recognition. *Cognition* 25, 1-2 (1987), 71–102.

[40] Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*. PMLR, 7076–7087.

[41] NYT. 2023. 'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead — nytimes.com. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html. [Accessed 02-May-2023].

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

[43] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language* 72 (2022), 101317.

[44] Alessandro Pianese, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. 2022. Deepfake audio detection by speaker verification. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–6.

[45] Ville Pulkki and Matti Karjalainen. 2015. *Communication acoustics: an introduction to speech, audio and psychoacoustics*. John Wiley & Sons.

[46] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220* (2019).

[47] Resemble AI. 2023. Introducing Neural Speech Watermarker - Resemble AI — resemble.ai. https://www.resemble.ai/neural-speech-watermarker/. [Accessed 16-08-2023].

[48] Brigitte Röder, Oliver Stock, Siegfried Bien, Helen Neville, and Frank Rösler. 2002. Speech processing activates visual cortex in congenitally blind humans. 16, 5 (sep 2002), 930–936. https://doi.org/10.1046/j.1460-9568.2002.02147.x

[49] Chunyin Shi, Luan Chen, Chengyou Wang, Xiao Zhou, and Zhiliang Qin. 2023. Review of Image Forensic Techniques Based on Deep Learning. *Mathematics* 11, 14 (2023), 3134.

[50] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), 5329–5333.

[51] Eugene C Snyder, Sanjana Mendu, S Shyam Sundar, and Saeed Abdullah. 2023. Busting the one-voice-fits-all myth: effects of similarity and customization of voice-assistant personality. *International Journal of Human-Computer Studies* 180 (2023), 103126.

[52] Joanna Stern. 2023. 24-Hour Challenge: Can My AI Voice and Video Clone Replace Me? — wsj.com. https://www.wsj.com/video/series/joanna-stern-personal-technology/24-hour-challenge-can-my-ai-voice-and-video-clone-replace-me/EC817295-03D0-4031-B40B-694D7BDE2797. [Accessed 03-May-2023].

[53] Synthesia. 2023. Ethical Deepfake Maker | Use Deepfakes for Good | Synthesia — synthesia.io. https://www.synthesia.io/tools/deepfake-video-maker. [Accessed 03-May-2023].

[54] Hugo Théoret, Lotfi Merabet, and Alvaro Pascual-Leone. 2004. Behavioral and neuroplastic changes in the blind: evidence for functionally relevant cross-modal interactions. 98, 1-3 (jan 2004), 221–233. https://doi.org/10.1016/j.jphysparis.2004.03.009

[55] Massimiliano Todisco, Héctor Delgado, and Nicholas W. D. Evans. 2016. A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients. In *Odyssey*.

[56] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. 2019. ASVspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441* (2019).

[57] Sue E Tranter and Douglas A Reynolds. 2006. An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing* 14, 5 (2006), 1557–1565.

[58] Xin Wang and Junichi Yamagishi. 2021. A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection. *Interspeech 2021* (2021).

[59] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. 2020. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language* 64 (2020), 101114.

[60] Robert Weeks, Barry Horwitz, Ali Aziz-Sultan, Biao Tian, C. Mark Wessinger, Leonardo G. Cohen, Mark Hallett, and Josef P. Rauschecker. 2000. A Positron Emission Tomographic Study of Auditory Localization in the Congenitally Blind. 20, 7 (apr 2000), 2664–2672. https://doi.org/10.1523/jneurosci.20-07-02664.2000

[61] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).

[62] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. 2015. Spoofing and countermeasures for speaker verification: A survey. *speech communication* 66 (2015), 130–153.

[63] Sung-Hyun Yoon, Min-Sung Koh, and Ha-Jin Yu. 2020. Phase Spectrum of Time-flipped Speech Signals for Robust Spoofing Detection. In *Odyssey*.

[64] Sung-Hyun Yoon and Ha-Jin Yu. 2020. Multiple Points Input For Convolutional Neural Networks in Replay Attack Detection. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), 6444–6448.

[65] Hong Yu, Zheng-Hua Tan, Yiming Zhang, Zhanyu Ma, and Jun Guo. 2017. DNN filter bank cepstral coefficients for spoofing detection. *Ieee Access* 5 (2017), 4779–4787.