Characterizing the Impact of Audio Deepfakes in the Presence of Cochlear Implant

Magdalena Pasternak, Kevin Warren, Daniel Olszewski, Susan Nittrouer, Patrick Traynor, and Kevin R. B. Butler
University of Florida, Gainesville FL 32611
{mpasternak, kwarren9413, dolszewski}@ufl.edu, snittrouer@phhp.ufl.edu, {traynor, butler}@ufl.edu

***Abstract*—Cochlear implants (CIs) allow deaf and hard-of-hearing individuals to use audio devices, such as phones or voice assistants. However, the advent of increasingly sophisticated synthetic audio (i.e., deepfakes) potentially threatens these users. Yet, this population's susceptibility to such attacks is unclear. In this paper, we perform the first study of the impact of audio deepfakes on CI populations. We examine the use of CI-simulated audio within deepfake detectors. Based on these results, we conduct a user study with 35 CI users and 87 hearing persons (HPs) to determine differences in how CI users perceive deepfake audio. We show that CI users can, similarly to HPs, identify text-to-speech generated deepfakes. Yet, they perform substantially worse for voice conversion deepfake generation algorithms, achieving only 67% correct audio classification. We also evaluate how detection models trained on a CI-simulated audio compare to CI users and investigate if they can effectively act as proxies for CI users. This work begins an investigation into the intersection between adversarial audio and CI users to identify and mitigate threats against this marginalized group.**

I. Introduction

Hearing loss is a significant global issue, impacting more than 1.5 billion people worldwide, as reported by the World Health Organization [69]. Cochlear implants (CIs) provide a valuable solution for individuals with severe hearing loss. These electronic devices stimulate the auditory nerve, allowing users to regain a modicum of hearing. With this technological advancement, CI users can increasingly benefit from audio-based modalities such as voice assistants [68].

However, rapid advances in audio technology and machine learning capabilities have given rise to new threats. *Audio deepfakes*, or synthetically generated audio, are based on deep learning algorithms and are convincingly realistic. They have been used to commit fraud [16], [60] and spread disinformation [47], [1]. A recent audio deepfake of US President Joe Biden [44] led the US Federal Communications Commission to outlaw robocalls that use voices synthetically generated by artificial intelligence [31]. While these scenarios underscore the threat of audio deepfakes to society, they do not consider the unique threats outside hearing persons (HPs) who are designed for *by default*. CI users, a group often marginalized even within the deaf and hard-of-hearing (DHH) community [55], may be more prone and susceptible to deepfake audio attacks.

Our goal is to determine the susceptibility of CI users to deepfake audio attacks. We also seek to inform the design of future defense systems against audio deepfakes based on the needs of this marginalized group. Therefore, in this paper, we aim to characterize the threats posed by audio deepfakes on CI users. We are interested in the potential differences in the challenges that audio deepfakes pose to HPs and CI users and in mitigating vulnerabilities disproportionately affecting the CI user community.

To understand CI users' perception of audio deepfakes and their detection capabilities, we model speech utterances and audio deepfakes using state-of-the-art CI simulators [48], [11], [8], [27], [10]. We also evaluate results from previous deepfake user studies [43], [34], [30], [37] and the performance of automated deepfake detectors [72], [59], [12], [5] against original and CI-simulated ASVspoof datasets. We then conduct a user study involving 35 CI users and 87 HPs, analyzing differences in responses in how each group differentiated natural from synthetic audio. Our analysis shows the need for a tailored defense, moving beyond a one-size-fits-all approach to deepfake detection. We hope that this work allows researchers to evaluate emerging audio deepfake threats without requiring significant imposition on CI users, who already face considerable obstacles to research study participation.

We thus make the following contributions:

- **Performance of Automated Deepfake Detectors** - We evaluate a corpus of synthetic audio using two well-established CI simulators [11], [8], [27], [10] and four automated deepfake detectors [19], [72], [5], [12]. Our analysis shows that deepfake detection on CI-simulated audio often performs similarly to the original, particularly with text-to-speech (TTS) methods, achieving an equal error rate (EER) of around 5%. We also show that the detection of CI-simulated audio is significantly worse when voice conversion (VC) approaches are used.
- **User Study of CI Users to Characterize Effects of Audio Deepfakes -** We present the first evaluation of the impact of audio deepfakes on the CI population with a user study of 35 CI users. We investigate the deepfake detection capabilities of participants, finding that CI users have a 70% detection accuracy. The study

Network and Distributed System Security (NDSS) Symposium 2025
24-28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.241117
www.ndss-symposium.org

also includes questions designed to identify what auditory features contribute to the user's perception of the audio authenticity. We compare their detection accuracy and guiding auditory cues across TTS and VC-generated deepfakes.

- **Comparative Study -** Additionally, we conduct a comparative study of deepfake audio perception between CI users and HPs, finding that HPs have a 7% higher detection rate across evaluated attacks. While it may seem intuitive that CI users are more susceptible to audio deepfakes than HPs, our findings surprisingly show that CI users can effectively identify specific TTS-generated deepfakes, yet are notably worse at detecting VC generated audio.
- **Evaluate the Use of Deepfake Detectors as Proxies for CI Users -** Finally, we show that CI users and deepfake detectors have similar difficulties detecting VC-generated audio, where the accuracy of human detection is 60% for the HP population and 48% for CI users. We compare results from detectors trained on breathing and high-level human-detectable feature with study participants, showing that their detection of TTS deepfakes is similar. Additionally, we extract the most influential detection regions from other ML deepfake detectors and compare them against CI users' survey responses. We find that when a speech sample has critical artifacts in higher frequencies (300-700 Hz), both ML detectors trained on CI audio and CI users struggle with correctly classifying the audio sample.

The remainder of this paper is organized as follows: Section II discusses the necessary background and related work; Section III provides our research goals; Section IV outlines the dataset used and methodology for modifying datasets with CI simulators and training and evaluating deepfake detectors; Section V discusses the procedures along with quantitative and qualitative results from our user study; Section VI compares user performance to ML detectors and assesses the feasibility of implementing a proxy for CI users; Section VII provides discussion and limitation; Section VIII concludes.

II. Background & Related Work

A. Speech Synthesis and Analysis

The emergence of Generative Adversarial Networks (GANs) has spurred critical advancements in generating persuasive "human-sounding" audio. While generative audio has many practical applications, including voice assistants (e.g., Siri, Alexa, and Cortana [26]), it also poses a security threat when used to imitate human speech [21], [54], [16]. In this context, computer-generated audio is called a "deepfake" [40].

Speech synthesis, widely known as text-to-speech (TTS), aims to generate naturally sounding human-like audio from written text. Despite various methods attempting to recreate natural and expressive voice [14], [36], [35], early versions of TTS lacked natural prosody, intonation, or emotional depth [54]. Another synthetic audio generation method, voice conversion (VC) [57], converts an actual human audio sample to match the target voice. This method allows the preservation of unique audio vocal traits and linguistic content, making VC-generated audio more challenging to detect [64].

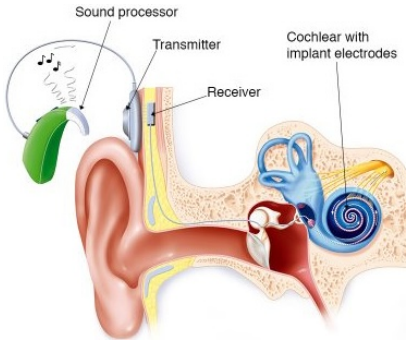


Figure 1: A typical CI system that converts captured sound to electric impulses delivered to the cochlea's hearing nerve [25].

Several ML deepfake audio detectors have been developed to differentiate human-generated and synthetic audio [1], [28], [29], [72], [17], [67]. They typically aim to identify artifacts absent in natural human audio or introduced during the deepfake generation process [1]. Some simple features found in many TTS-created deepfakes, such as the presence of silence or variations in rhythm, intonation, and style, can help the ML model learn shortcuts to discriminate deepfakes [41], [42], [65], [3], [12]. An alternative approach by Blue et al. reconstructs the physical characteristics of a speaker's vocal tract from their voice recordings [5]. The detector flags audio as synthetic if estimated vocal tract structures are unlikely to belong to a human [5].

With the substantial progress demonstrated in deepfake detection capabilities and improving deepfake qualities, recent research has sought to assess the abilities of humans to detect deepfakes and compare human performance with automated deepfake detectors [43], [34], [30], [37]. However, these efforts focus on the general population without considering potentially more susceptible groups. Therefore, we explore the effects of audio deepfakes on CI users, whose needs have yet to be addressed in deepfake detection literature.

B. Cochlear Implant

A CI is an electronic device surgically implanted under the skin that directly stimulates the auditory nerve, restoring functional hearing [71], [45]. Figure 1 shows the external sound processor behind the ear, which uses a built-in microphone to pick up various sounds, such as speech, and translate them into digital signals. Subsequently, these signals are transformed and encoded into radio frequency (RF) signals, which are then sent to an antenna inside the transmitter [71], [39], [25]. The receiver, placed under the skin behind the ear, decodes the RF signal, converts it into electric currents, and directs it into the cochlea [71], [39], [25]. Implant electrodes bypass damaged cochlear hair cells and directly stimulate the auditory nerve

with an electric current. The resulting electrical impulses in the auditory nerve are perceived as sound.

While many CI users report notable enhancements in quality of life, they continue to face challenges in daily activities, such as phone communication [53]. These challenges stem from the inability, or reduced capability, to detect pitch changes and variations in speech prosody [15]. This results in difficulties distinguishing between statements and questions or identifying emotions conveyed through speech [15]. Additionally, limited spectral information [18] and problems processing auditory information over time affect the ability of CI users to perceive rhythm, which further contributes to challenges involving audio interfaces [46].

C. CI Simulation

CI users differ across various demographic and individual factors, including age, sex, age at CI implantation, duration and degree of deafness, hearing loss, auditory system health, and remaining cochlear and neural function [71], [45], [53], [56]. This variability, coupled with the inability to isolate specific variables, the numerous choices in signal processing, and the challenge of recruiting participants from relatively low-incidence populations, leads to highly variable response data in research involving CI users. Research becomes time-consuming and expensive, underscoring the need for better approaches and tools to control and investigate parameters affecting CI user speech perception.

Therefore, reliable proxy models have become a crucial tool in CI research [11], [8], [27], [10], [22]. Researchers have developed several CI simulators to understand CI user perception of speech and sound. One of these approaches involves noise-vocoded speech, obtained by partitioning speech signals into separate, logarithmically spaced frequency bands. The amplitude envelope is extracted from each channel and modulates noise in that frequency band. Afterward, these bands are recombined to create noise-vocoded audio [56]. In the preliminary phase of assessing novel sound-processing algorithms for CIs, vocoders play a critical role by facilitating the evaluation of these algorithms with HPs. Findings from these evaluations have shown a significant correlation with the results of experiments conducted on high-performing CI recipients [61]. Simulators are valuable tools for evaluating device-related challenges in speech perception and assessing CI capabilities and limitations, streamlining otherwise time-consuming and costly research processes [61], [4], [13], [56]. Our research explores the practicality of using CI simulation within the cybersecurity domain.

III. Research Goals

The rise in both the prevalence and sophistication of audio deepfakes and a surge in malicious activities they facilitate, including fraud and misinformation, represent a security threat [16], [47]. CI users, a potentially more vulnerable population, are not represented within user profiles examining the effects of audio deepfakes. Consequently, this study aims to investigate CI users' perception of audio deepfakes, measure their susceptibility, and explore their detection strategies.

To understand CI users' potential susceptibility to audio deepfakes, we encode deepfake audio with CI simulators and evaluate them against ML-based deepfake detectors. Subsequently, we conduct a user study to compare how CI users detect audio depfakes with HPs. Finally, we evaluate whether future research on CI users' deepfake detection can be streamlined using ML deepfake detectors as proxies.

Thus, we aim to explore the following research questions:

- **RQ1.** Are there differences in the performance of automated deepfake detectors on original and CI-simulated audio?
- **RQ2.** Are CI users more susceptible to deepfake audio attacks than HPs?
- **RQ3.** How do CI-simulated detection models compare to actual CI users, and can these models serve as adequate substitutes?

IV. CI Simulation and Automated Detection

We begin by processing audio deepfake samples using CI simulators to simulate how CI users perceive this audio. We seek to determine whether CI users are susceptible to all audio deepfake attacks or whether particular types of attacks garner additional concerns.

A. Dataset Selection

The ASVspoof challenge has emerged as the premier anti-spoofing competition against deepfakes, introducing new iterations since its 2015 inception. The most recent edition, ASVspoof2021, asks participants to design countermeasures that shield automatic speaker verification (ASV) systems from voice spoofing attacks, similar to previous versions [70]. This dataset includes logical access (LA), deepfake task (DF), and physical access (PA) scenarios. The evaluation dataset amalgamates audio samples from previous years with many new-generation methods employing over 100 spoofing algorithms. The LA and DF tasks in the challenge concentrate on lossy compressed audio, considering telephony encoding and transmission in the audio generation process.

As our research emphasizes the differences between deepfake audio attacks, we also include the ASVspoof2019 dataset, which focuses on spoofing attacks executed using TTS and VC techniques [62]. This dataset covers 122,157 samples (43% male and 57% female), featuring samples from native speakers and 19 different audio generation techniques [38]. Table A1 in the Appendix A presents a comprehensive overview of these attacks and their generation methods. Regarding distribution, the training and development sets contain samples from techniques A01 to A06 (where A01-A04 are TTS deepfakes and A05-A06 are VC). In contrast, the evaluation set includes samples from techniques A07 to A19 (A07-A12 and A16 are TTS, A17-A19 are VC, and A13-A15 are a mix of both) [66].

B. CI Simulators

We use two distinct vocoder-based CI simulation tools to mimic the auditory perception of CI users listening to deepfake audio. The audio samples, described in Section IV-A, were processed through both simulators.

Generic MATLAB Toolbox (GMT): The Generic MATLAB Toolbox (GMT) is a collection of scripts designed to simulate the Advanced Bionics (AB) CI [67] emulating the current HiRes 120 processing strategy. It incorporates a noise vocoder CI simulation, leveraging an FFT to model the cross-talk between adjacent channels of the CI[1]. This toolbox has been used to understand spectral modulation and speech and music perception, create enjoyable music for CI users, and predict neural responses to speech stimuli [7], [63], [22], [2].

MATLAB Vocoder: A MATLAB script uses the noise-vocoding method to emulate the audio effects of a CI [48]. It is a popular tool used in CI research [48], [11], [8], [27], [10]. The audio undergoes band-pass filtering, featuring a low-frequency cutoff at 0.25 kHz and a high-frequency cutoff at 8 kHz. Following practices in CI research, we opted for a noise vocoder that employs five channels, establishing boundaries at 0.6 kHz, 1.2 kHz, 2.3 kHz, and 4.3 kHz.

C. Model Selection

We compare the original and CI-simulated datasets across distinct machine learning-based detectors: FastAudio, AIR-ASVspoof, BTS-E, and vocal tract reconstruction (VTR) [72], [59], [12], [5]. We evaluate their overall detection rates and analyze the outcomes of each spoofing attack to determine if specific techniques pose more significant challenges for CI users. Additionally, we use VTR to distinguish between original and CI-simulated samples [5].

AIR-ASVspoof: Zhang et al. introduced a voice spoofing detection system that detects unknown voice spoofing attacks using one-class learning [73], [72]. This detection model is based on the ResNet-18 framework, whose deep architecture enables it to capture and learn low-level features (e.g., simple frequency patterns) and high-level features (e.g., complex temporal changes). The authors developed a one-class softmax (OC-Softmax) function incorporating attentive temporal pooling, outperforming most single-system classifiers without data augmentation techniques [24].

FastAudio: FastAudio was designed to select features from an audio dataset intended for processing by a detector that determines whether the audio is synthetic. This preprocessing aims to preserve the maximum amount of relevant features while minimizing file size. We selected FastAudio because it was a top-performing entry in the ASVspoof 2019 challenge and demonstrated strong performance in the 2021 challenge [59], [19].

Breathing-Talking-Silence Encoder: The BTS-E framework detects deepfakes by analyzing the correlation between breathing, speech, and silence within an audio sample [12]. The authors argue that synthetically generating natural human sounds, such as breathing, poses significant challenges for text-to-speech (TTS) deepfakes, making it a valuable feature for detection. Unlike other methods, we exclusively trained and evaluated the BTS-E model on TTS-generated audio using the most effective variation: a transformer encoder with a concatenation fusion strategy and 32 output channels.

[1]https://github.com/jabeim/AB-Generic-Python-Toolbox

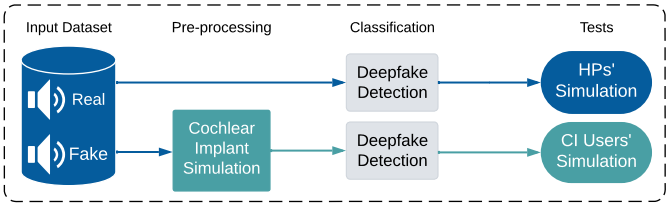


Figure 2: Detection training and evaluation pipelines for original and CI-simulated audio representing each testing scenario.

Vocal Tract Reconstruction (VTR): Blue et al. employed a fluid dynamic technique to estimate the physical parameters of a vocal tract from a voice recording [5]. The authors discovered that deepfake audio frequently produced unrealistic vocal tracts, prompting them to develop a deepfake audio detector based on this observation. This detector isolates speech segments during data extraction, ideally distinguishing between fake and real audio samples. Subsequently, these optimal speech segments differentiate the two audio types during an evaluation phase. In our research, we executed the data extraction and evaluation phases for this detector in each test case, consistent with the original authors' design.

D. Testing

Metrics: We assess the effectiveness of different synthetic audio detectors by employing metrics such as recall, precision, and Equal Error Rate (EER). *Precision* indicates the proportion of correctly identified positives to all positives, whereas *recall* measures the number of labeled attacks that are actual attacks. The EER metric measures the performance of the anti-spoofing systems by analyzing the countermeasure (CM) score. CM score reflects how closely a given speech sample resembles actual speech. To determine the EER, the algorithm establishes threshold values at the point where the rates of false positives and false negatives are equal. A low EER value signifies a high level of detector reliability, making it a key benchmark in the ASVspoof challenges [62].

We comprehensively evaluate similarities and differences between the original and CI-simulated audio by following the pipeline process shown in Figure 2. The first test (HP simulation) establishes our baseline and aims to mimic the performance of HPs. In the second test(CI user simulation), we preprocess the audio dataset using CI simulators and then train and evaluate deepfake detectors on these preprocessed files.

For the VTR detector, we conduct four tests on combinations of real and deepfake audio samples across original and CI-simulated datasets. Given the high computational demand of the detection via vocal tract reconstruction, we randomly

Table I: EERs for ASVspoof2019 and 2021 challenge across two detectors tested on original and CI-simulated (GMT, MVC) audio. Overall, detectors show performance degradation when evaluated on CI-simulated files (Test 2).

Dataset	Detector	Test 1	Test 2	
			GMT	MVC
ASVspoof2019	AIR-ASV	2.84%	9.42%	9.25%
	FastAudio	5.36%	8.42%	8.82%
ASVspoof2021	AIR-ASV	9.54%	17.26%	16.42%
	FastAudio	11.56%	15.37%	12.67%

sample between 80-100 audio files from each attack approach, resulting in 2,300 actual and deepfake audio samples. We compute across the previously described tests to detect differences in the approximate vocal tract configurations and core fluid dynamics between datasets. Unlike other detectors, VTR uses precision and recall to assess results. These two metrics allow the evaluation of the two aspects of the model's accuracy, helping to prioritize the potential victims from deepfake audio attacks rather than minimizing false alarms on genuine audio. Finally, we conduct two more experiments to determine whether this detector is sensitive to the CI simulation effects.

E. Deepfake Detectors Results and Analysis

Here, we state a summary of the key findings as follows:

1) Deepfake detection algorithms face greater challenges when identifying deepfakes in CI-simulated audio.
2) VC deepfakes result in a higher EER in deepfake detection tasks than TTS deepfakes.
3) When detectors are tasked with identifying a specific feature synthesized in TTS deepfakes, such as breathing, the detection performance among HPs and CI users is similar. However, this detection method is effective only for TTS deepfakes.
4) The BTS-E detector, which extracts explicit audio features using breathing sounds, can detect the original and CI-simulated audio TTS deepfakes equally well.
5) The effectiveness of deepfake detection through vocal tract reconstruction is comparable to original and CI-simulated audio as the simulation does not affect underlying vocal tract properties.

Detection Results: AIR-ASVspoof, achieves a 2.84% EER on the original ASVspoof2019 files for test 1 and when trained and evaluated on the CI-simulated dataset (Test 2), its EER rose to 9.42% with the GMT CI simulator and 9.25% using the MVC CI simulator. Similarly, the FastAudio deepfake detector has a baseline EER of 5.36%; in Test 2, EER gets 8.42% for GMT and 8.82% for MVC CI-Simulator.

The FastAudio synthetic audio detector achieves the smallest difference between Test 1 (original audio) and Test 2 (CI-simulated audio). The difference is 3.06% and 3.46% for the GMT and MVC CI simulators, respectively, on the ASVspoof2019 dataset. For the ASVspoof2021 data, the slightest difference is when using the FastAudio synthetic audio detector, with a 16.09% difference for both CI simulators. This demonstrates that FastAudio learns CI-simulated features

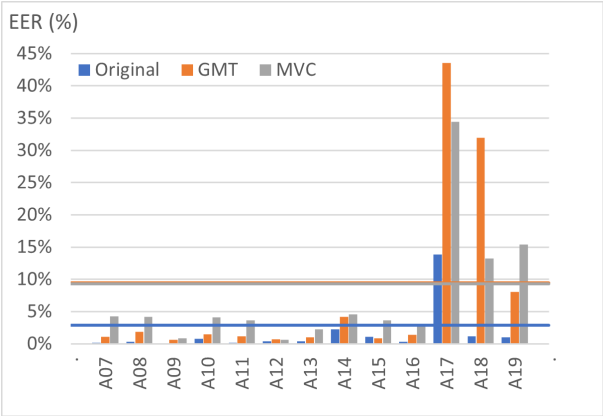


Figure 3: EERs for AIR-ASVspoof detector on original and CI-simulated datasets using GMT and MVC simulators. EER for TTS deepfakes on all three datasets is low, while the EER for VC attacks (A17-A19) significantly increases when simulated.

better than AIR-ASVspoof and generalizes CI-simulated audio more effectively.

This trend is also observed in the ASVspoof2021 dataset. The AIR-ASVspoof deepfake detector has an EER of 9.54% on the original 2021 dataset. The EER increases to 17.26% when run on the dataset modified by the GMT simulator and 16.42% for the MVC CI simulator. The more consistent performance of FastAudio, increasing from 9.54% for Test 1 to 15.37% (GMT) and 12.67% (MVC), suggests that its design, which prioritizes capturing the peak frequencies around first and second formants, contributes to its ability to generalize when anticipated audio characteristics are blurred or absent.

Unlike previously evaluated deepfake detectors, BTS-E bases detection on breathing sounds, which is helpful when this feature is entirely synthetically generated (e.g., in TTS deepfakes). As represented in Figure 5, the average EER across all attacks (including VC audio) is 16.4% for original audio, slightly increasing to 17.6% and 16.6% when subjected to GMT and MVC simulators, respectively. However, when evaluated only on TTS deepfakes, the EER of the original dataset is 0.394%, and CI-simulated EER reaches 0.326% and 0.491% for GMT and MVC, respectively.

Comparison between deepfake generation methods: We evaluate all deepfake generation approaches (Table A1) included in the ASVspoof2019 dataset. As represented in Figures 3 and 4, the most challenging attack on the original dataset to discern for both deepfake detectors is VC waveform filtering (A17). The EER of the AIR-ASVspoof detector rises significantly to 13.82%, and for FastAudio, the EER goes up to 13.48%. Another complex attack for FastAudio to differentiate real from fake audio is A18, resulting in 22.84% EER.

The varied EERs observed with different deepfake generation techniques indicate these techniques' inherent complexity and variability. The VC deepfake generation approach is incredibly challenging for both detectors [62]. This might be

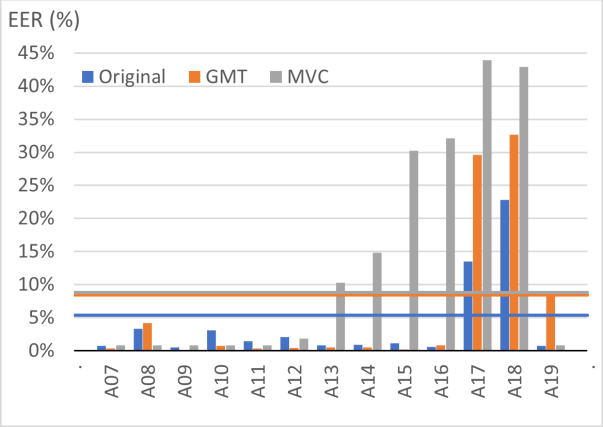


Figure 4: EERs for FastAudio detector on original and CI-simulated datasets using GMT and MVC simulators. CI users demonstrate a better EER for TTS than the HPs group while continuing to have large EER increases on VC attacks.

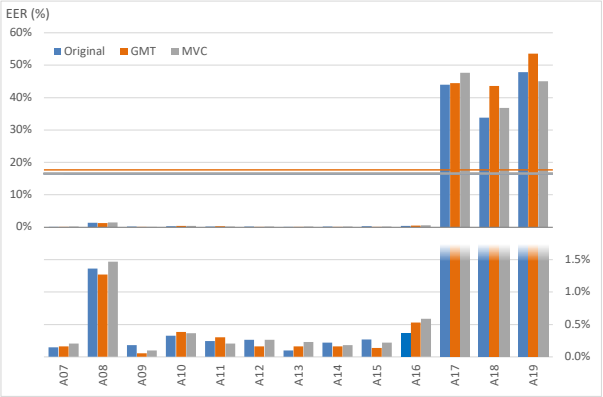


Figure 5: EER rates for the BTS-E detector on original and CI-simulated datasets using GMT and MVC simulators. The panel chart, with a top panel showing all the data (left y-axis) and a bottom panel highlighting the first 1.5% of EER (right y-axis). It demonstrates that all groups have low EER for TTS deepfakes (A07-A16) and high EER for VC deepfakes (A17-A19).

due to the generation technique of VC, which introduces subtle nuances that closely mimic actual audio features, making differentiation challenging. For both CI-simulated datasets, detectors struggle with correctly classifying VC deepfake attacks (A17-A19).

AIR-ASVspoof and FastAudio detectors achieve relatively low EER on the original dataset, but their performance deteriorates significantly under CI simulation, whereas the BTS-E detector has a minimal increase. Its design focuses on inconsistencies in breathing, talking, and silence, which are particularly effective. This suggests that relying on unnatural breathing, talking, or silence features may offer an advantage in TTS-generated deepfake detection.

When focusing on TTS-based attacks (A07-A16), the BTS-E detector performs similarly across original files and simulators, with only minor fluctuations in EER. For specific attacks such as the A09 (vocoder generation), there is a difference in performance between original and CI-simulated files, as the EER drops under simulated conditions. This suggests that such simulation may, in some cases, make detection easier or more complex depending on the nature of the attack. For VC deepfakes, BTS-E detection EER rises significantly, reaching 47.6% for A17 using the MVC CI-simulator. As the VC systems convert original audio to sound like a target speaker and use non-speech frames in the training phase, the detector cannot distinguish between real and VC-generated deepfake audio, as anticipated by the authors [12].

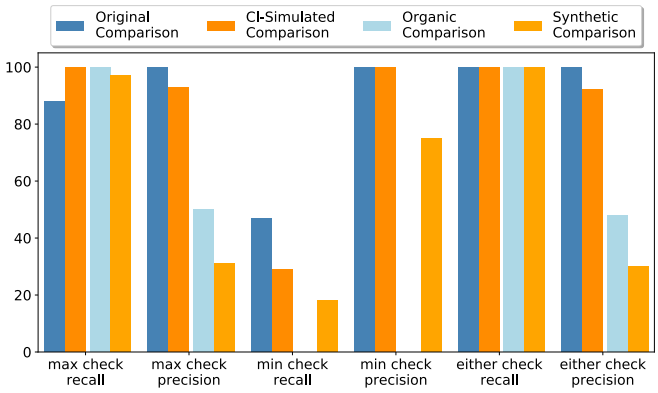


Figure 6: Comparison of max and min checks for recall and precision evaluated with detector relying on vocal tract reconstruction.

AIR-ASVspoof and FastAudio detectors achieve relatively low EER on the original dataset, but their performance deteriorates significantly under CI simulation, whereas the BTS-E detector has a minimal increase. Its design focuses on inconsistencies in breathing, talking, and silence, which are particularly effective. This suggests that relying on unnatural breathing, talking, or silence features may offer an advantage in TTS-generated deepfake detection.

Detection by Vocal Tract Reconstruction: The detection of the vocal tract reconstruction detector attains 100% precision on the original real and fake audio datasets. When testing the CI-simulated dataset, this detection method can differentiate between real and deepfake audio with a precision and recall of 100% and 93%, respectively. The drop in the recall comes from the increased noise and overall lower audio quality caused by the CI simulator.

The relatively unchanged performance of the detector indicates that the core fluid dynamics principles in the speech remain intact despite the additional noise and distortion caused by the CI simulator. Although this method yields a high precision-recall score, the system makes a trade-off with a relatively high false positive rate. This could potentially lead to significant numbers of false alarms, subsequently increasing the threat-alert fatigue of users and reducing reliance caused by mistranscription during the preprocessing stage.

Figure 6 shows the comparison between the original organic

audio with the CI-simulated audio and the original synthetic audio with the simulated synthetic audio. The detector could differentiate the non-simulated and simulated organic audio with a precision and recall of 100% and 50%, respectively. Similarly, the detector differentiated the non-simulated and simulated synthetic audio with a precision and recall of 100% and 30%, respectively. In both cases, the detector's overall performance drops considerably, highlighting that the detector is struggling to differentiate the non-simulated and simulated samples. This further indicates that the CI simulator's underlying vocal tract properties leveraged by the Who Are You detector are unaffected.

The performance of detectors declines when evaluated on CI-simulated audio compared to the original audio, which addresses the **RQ1**. This is especially apparent for AIR-ASVspoof and FastAudio detectors, where the EER on CI-simulated audio is elevated, especially on VC deepfakes, which include natural human audio features. The FastAudio, by evaluating peak frequencies, shows some adaptability to CI-simulated audio. BTS-E remains relatively robust for TTS deepfakes because it relies on anomalies synthetically generated by breathing and silence cues. Similarly, vocal tract reconstruction-based detections result in high precision scores, as the core fluid dynamic properties are preserved during CI simulation.

V. User Study

We conduct a user study to characterize the detection capabilities of CI users. Our approach involves assessing participants' ability to distinguish whether provided audio samples are generated by humans or synthetically. We also gather participant insights on the rationale behind their decisions, including cues they use to determine audio authenticity. We then perform a comparative analysis between CI users and HPs to identify any differences in the detection between these groups. We structure our survey around the central null hypothesis as follows:

H_0: There is no difference in deepfake detection between HPs and CI users.

We aim to identify variations in detection approaches, key auditory cues used in detection, and other factors that influence the accuracy of audio detection.

A. Dataset

Based on results from automated deepfake detectors, we consider deepfake generation techniques where detectors do poorly or well based on EER scores (e.g., as shown in Figures 3-5). We base our experiments on the evaluation split of the ASVspoof2019 dataset and examine five deepfake generation approaches to generating audio deepfakes. Two techniques employ TTS synthesis models (neural waveform-A08 and vocoder-A09), and three use VC methods for deepfake creation (waveform filtering-A17, vocode-A18, and spectral filtering-A19). We aim to examine deepfake generation approaches, TTS, and VC to assess how they impact the CI population. Due to the limited number of CI participants, we selected a pool of 660 samples and used stratified sampling to choose equal numbers of real and deepfake audio samples. Within the deepfake group, we choose an equal amount of audio samples from each of the five spoofing systems. The selection of 20 audio samples for each participant is randomized from a total pool of samples for each group, where each audio sample is repeated three times for HPs and two times for the CI group. We presented the same speech samples to HPs and CI users to compare the group results.

B. Ethical Considerations

Before working with human subjects, we acquired approval from our Institutional Review Board (IRB). Each survey began with a Participant Informed Consent page, in which we outlined the survey's purpose and procedure. We informed participants that they could withdraw from the study at any point. We described the data collection process, highlighting that any collected data will be deidentified and anonymized. We then asked users to proceed only if they consent to participate in the study. Additionally, to ensure compliance with European and UK laws, we provide a separate consent form under the General Data Protection Regulation (GDPR) supplement.

We also engaged with researchers experienced in working with d/Deaf and Hard of Hearing (DHH) communities to ensure our research aligns with ethical standards and cultural norms. This included respecting the choices of individuals who opt for CIs while also acknowledging and valuing the perspective of those within the d/Deaf community opposed to CIs [20], [9].

C. Recruitment and Experiment

We use Prolific [52], a crowd-sourcing platform, to facilitate the recruitment and management of participants for our study. The quality of data acquired through Prolific is comparable to that from the dominant crowd-sourcing platform, Amazon Mechanical Turk [50]. Using Prolific's filters, we selectively target our study to users aged 18 or above, with English as a primary language. Additionally, for our study group, we filter participants for those who use CIs, which yields fewer than 100 participants for the CI group. Consequently, as a secondary recruitment method, we advertise our study in a Facebook group focused on CI users.

Experiment: We conduct our study through a website form, beginning with introducing participants to the study goal and methodology. Each participant is presented with 20 randomized audio samples from our dataset, consisting of real and deepfake audio files. We instruct participants to listen to each audio file, determine whether it is human (Real) or computer-generated (Fake), and rate their decision confidence on a 5-point Likert scale. Additionally, we ask participants to explain what factors had influenced their decision. We monitor how often each audio sample is played and the time spent on each question to assess their effects on the detection task. Participants are required to provide all responses before they

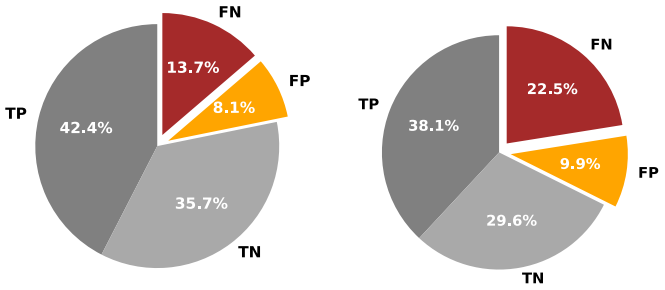
can proceed to the next question. To prevent bias from question sequencing, we disable participants from returning to previous questions to change their answers. We do not provide any feedback to participants regarding their performance, which mimics a more authentic situation and avoids influencing their future decisions. Within the survey, we included two attention checks, allowing us to exclude data from participants who were not paying full attention. Lastly, to better understand the diversity of our participant pool, we collect demographic information, including gender, age, the highest level of education, the presence of hearing loss, the type and fit of hearing aids, number of years since CI implantation, and length of daily CI usage.

Participants: Over nine months, we received feedback from 215 participants across both groups. Despite many participants initially claiming to use a CI on the Prolific website, further verification through our questionnaire revealed that several of them had not received an implant. According to Prolific guidelines [52], we followed up with participants who provided inconsistent responses between our survey and Prolific's filters to clarify their answers. We compensated all participants for their time, but within our evaluation, we included only those participants who confirmed using a CI, either through Prolific filters and directly in the web survey or after resolving any discrepancies between their responses. After filtering out participants who did not meet our research criteria, failed to complete the study, or did not pass attention checks, we finalized a cohort of 1,740 responses from 87 HPs and 700 responses from 35 CI users.

All participants are 18 or older, most in the 25-34 age group. The gender distribution is 58% male, 40% female, and 2% non-binary. The most common educational level was an undergraduate degree, with a minority of participants having attained additional higher education. Among CI users, 61% have a unilateral CI fitting, while 39% reported using either bilateral fitting or CI and hearing aids. On average, participants have had their CI for 7 years, and 92% use their implants for more than 7 hours daily.

D. User Data Results and Analysis

Metrics: To evaluate human performance on the dataset, we use three primary metrics: precision, recall, and accuracy. *Precision* indicates the proportion of samples labeled by participants as computer-generated that were indeed deepfakes. *Recall* measures the ratio of correctly identified deepfakes to the total number of deepfake speech samples. *Accuracy* represents the ratio of correctly identified audio samples to the total number of speech samples provided to participants. To determine the statistical significance of our results, we employ an independent t-test, which compares the means of two groups to ascertain if they are significantly different. For comparisons involving multiple groups, we utilize an ANOVA test. We use the Pearson test to assess the likelihood of a causal relationship between the two variables and set the threshold for statistical significance at $p < 0.05$ [51].



(a) HPs group confusion matrix (b) CI users confusion matrix

Figure 7: Depiction of detection rates between (a)HPs and (b)CI users. Both groups detected a comparable number of audio files as real, 56.1% (11.23 files) for HPs and 60.6% (12.12 files) for CI users, yet HPs correctly classify 1.4% more real and 6% more fake speech samples compared to CI users.

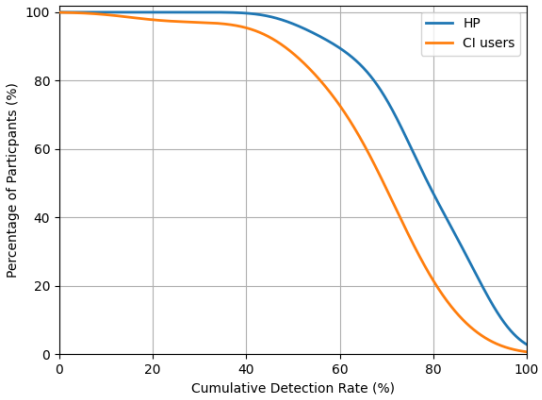


Figure 8: Cumulative distribution function for participant accuracy at or below detection rate, i.e., a detection rate of 20% represents the percentage of study participants who accurately detect at least 20% of audio samples. The detection rate is generally at least 9% lower for the CI users group than for HPs, most of whom can detect deepfakes with 77.5% accuracy or higher.

Overall Performance: Within our study, HPs detected deepfake audio with 78.1% detection accuracy, whereas CI users performed significantly worse ($p = 0.0002$), achieving only 67.6%. As illustrated in Figure 7, this difference is primarily due to the higher false-negative rate among the CI users, who could identify only 57% of deepfakes successfully. They inaccurately labeled 22.5% of deepfake audio, which resulted, on average, in 12.12 audio clips labeled as real, while only ten human-generated audio samples were given. In contrast, HPs marked, on average, 11.23 files as real. This difference points to the challenges CI users might face during audio authentication tasks.

The cumulative distribution function plot in Figure 8 shows the percentage of participants capable of detecting deepfakes at

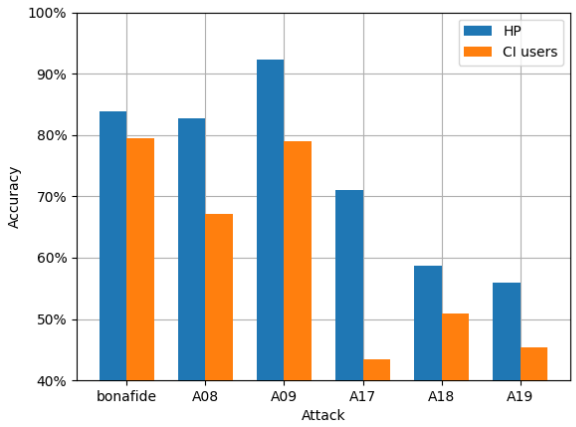


Figure 9: Detection rate across real audio (bonafide) and TTS (A08 & A09) and VC (A17-A19) deepfake generation methods for HPs and CI users. HPs have a higher detection rate than CI users for each attack. Both groups have more difficulties discerning audio generated by VC models than TTS.

a given detection rate or higher. Indicated by an initial steeper slope, the detection distribution for CI users leans towards lower accuracy ranges, where many of them cannot correctly classify 50% of audio samples. The disparity emerging at the 50% detection mark shows that *HPs consistently surpass CI users in deepfake detection* by approximately 9%. This demonstrates a significant and systematic issue when distinguishing between authentic and deepfake audio across the entire CI user population.

Deepfake Generation Method Comparison: While assessing users' detection accuracy on VC-based deepfakes, we observed a notable decrease in detection performance across both groups. HP's accuracy on the A17 attack was the highest among VC-based deepfakes (71.1%), while CI users had accuracy below the random guessing rate at 43.4% (Figure 9). The A18 and A19 attacks, which consistently fooled participants in both groups, are particularly concerning as they resulted in accuracy rates similar to random guessing. Notably, HPs significantly ($p_{VC} = 1.186 \times 10^{-4}$) outperformed CI users across all three VC techniques. The lower detection rates in the CI group indicate increased susceptibility across all groups of attacks.

In TTS-generated deepfakes, HPs misclassified only 13.7% of the fake audio, while CI users misclassified 24.8% of fake audio, nearly doubling the error rate. For VC-generated deepfakes, the gap was narrower, but the overall challenge of the detection was greater across both groups: HPs misclassified 41.2% of the fake audio as real, and CI users incorrectly labeled a majority of deepfakes (52%) as real. These findings highlight the nuanced difficulties both groups face with different deepfakes types, suggesting that they, despite their awareness of deepfakes, are more easily deceived by the naturalness of VC-generated audio. CI users, on the other hand, struggle significantly across both categories, indicating the heightened vulnerability to audio manipulation.

This disparity stems from the fundamental differences in how these technologies generate audio and how cochlear implants process sound. Cochlear implants rely on electrical stimulation to simulate auditory perception, often reducing the fidelity of subtle speech cues such as pitch variation, prosody, and harmonic richness. VC technology, which uses a real human speech sample (the source) and transforms it to match the vocal characteristics of another speaker (the target), retains many of these natural elements, making it particularly challenging for CI users to detect manipulations. In contrast, TTS deepfakes synthesize speech entirely and are often easier to identify due to inherent flaws such as unnatural prosody, robotic intonations, and a lack of emotional depth. Additionally, TTS-generated speech lacks human subtleties like breathing, pauses, hesitations, and irregularities in pitch or volume - many of which CI users may use to guide their detection.

These findings highlight the need for targeted interventions to support CI users in detecting audio deepfakes. Strategies could include developing assistive tools that amplify or visualize subtle inconsistencies in speech or incorporating deepfake detection training into CI rehabilitation programs. By addressing these perceptual challenges, we can help CI users navigate an increasingly complex audio landscape where authentic and manipulated speech is becoming harder to distinguish.

Audio Features Influencing Participant Judgments: In the survey, participants were asked to classify audio as fake or real and to specify the auditory cues influencing their judgments. To analyze this qualitative data, we employed thematic analysis methods outlined by Braun and Clarke [6]. Following familiarization with the data, we generated initial codes based on notable features across participant responses through multiple rounds of analysis. Two independent reviewers evaluated each response, assigning relevant subtheme keywords or marking it as irrelevant. These keywords were subsequently organized into subthemes, broader themes, and major categories as detailed in Appendix C in Table C1.

The analysis (Figure 10) reveals distinct challenges HPs and CI users face when detecting TTS and VC-generated deepfake audio. Both groups relied on different auditory features to varying extents, indicating a complex landscape of perceptual factors influencing detection strategies.

Nearly half of the auditory cues provided by both HPs and CI users (46% for HPs and 44% for CI users) were related to speech prosody. These cues resulted in 75% accuracy for HPs and 68% for CI users in identifying deepfakes. The lower accuracy for CI users suggests that they struggle with nuanced features such as rhythm, intonation, and stress patterns, which are often not fully conveyed by CI processing. Emotion detection was particularly challenging for CI users, who achieved only a 60% accuracy compared to higher rates for HPs. These findings align with existing research, as discussed in Section II-B, indicating that CI users

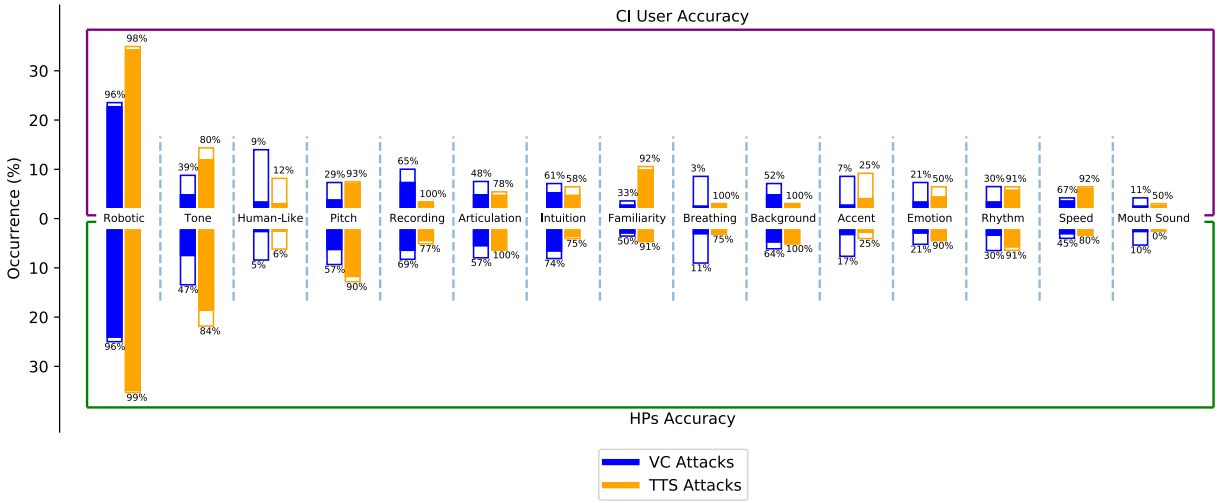


Figure 10: This figure shows the performance of CI users and HPs on TTS and VC attacks arranged by theme and its proportion (y-axis). The y-axis shows the proportion of the theme (i.e., the larger the bar, the more it appeared in our study). It then shows the accuracy as a filled bar.

often have difficulty perceiving emotional tone and prosodic features. Such limitations highlight a key vulnerability in CI users' ability to discern manipulations that rely on naturalistic emotional expression.

Interestingly, while both groups performed similarly when relying on the accent as a cue (between 68-70%), CI users were more reliant on this feature (10.5% vs.6.3% for HPs). This suggests that CI users prioritize more prominent auditory cues, whereas HPs adopt a more distributed approach, likely due to their ability to access a broader range of acoustic signals.

The most reliable cue for CI users was robotic-like sounds, with accuracy rates of 97.9% for TTS and 96.1% for VC deepfakes. This aligns with a general familiarity with synthetic speech, such as voice assistants, which often exhibit slightly robotic characteristics. Interestingly, while CI users performed well on real audio while relying on human-like features (96.3% accuracy), they frequently misattributed these features to VC deepfakes (12.5% of cues). HPs followed a similar pattern but were less prone to this error (6.2%), suggesting that CI users may overgeneralize certain cues when distinguishing real and fake audio.

A significant gap emerged in using pitch and inflection to identify VC deepfakes. While pitch was a reliable cue for both groups in detecting real and TTS-generated audio (82-92% accuracy), CI users performed poorly on VC deepfakes, achieving only 29.2% accuracy, compared to 56.5% for HPs. This further suggests that CI users struggle to extract subtle pitch and inflection patterns from VC audio, which retains much of the natural variation of real speech.

Speed proved to be a useful cue for TTS deepfakes, with CI achieving 91.7% accuracy compared to 45.5% for VC-generated audio. Notably, CI users relied more on speed as a cue (4.4% of cases) than HPs (1.1%), suggesting that CI users focus more on audio features less affected by CI processing. However, their reliance on speed cues for VC deepfakes, which often mimic natural speech rhythms, led to decreased accuracy.

Technical aspects such as *sound quality* and *vocal artifacts* (e.g., breathing sounds, lip smacks, tongue movements) also revealed important differences. While HPs can rely on background noises as a clue, CI users' accuracy decreased by 17.9 percentage points when relying on these cues. Additionally, *vocal artifacts* were detected less accurately by CI users, indicating these cues are less effective for distinguishing between deepfake and authentic audio. This finding suggests a potential misunderstanding among CI users regarding the ability of deepfake technology to replicate natural imperfections.

Breathing patterns emerged as a misleading cue for both groups when identifying VC deepfakes. HPs correctly identified 96% of real audio and 75% of TTS deepfakes but achieved only 11.4% accuracy for VC samples when relying on breathing. CI users performed even worse, with a detection rate of 3.3% detection rate. This reflects the sophistication of VC models in preserving the natural breathing patterns, which makes them particularly deceptive. Conversely, the more artificial breathing patterns of TTS deepfakes made them easier to detect (91% accuracy for HPs, 92% for CI users).

In summary, the findings provide critical insights into differences between HPs and CI users in detecting deepfake audio, addressing **RQ2**. CI users face heightened challenges in detecting deepfake audio, particularly VC-generated samples, and rely on prominent cues that suggest a strategic adaptation to their perceptual limitations. As the approaches used to detect deepfakes differ among the groups, including overall performance, detection strategies, and reliance on different acoustic features, we can reject our null hypothesis and state that there are differences in deepfake detection between HPs and CI users.

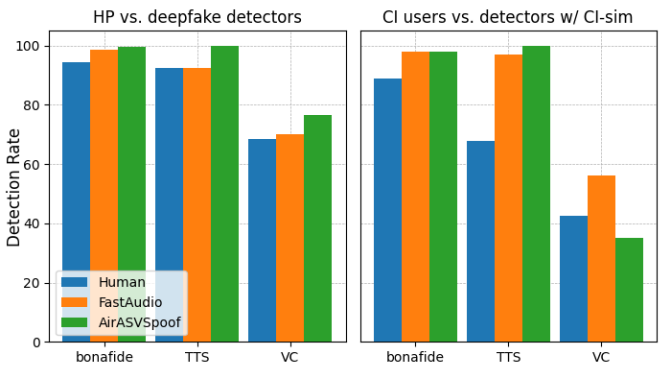


Figure 11: Comparison of detection accuracy across study participants and AIR-ASVspoof and FastAudio automated deepfake detectors across real and deepfake (TTS & VC) audio. The left graph displays rates for HPs and deepfake detectors trained and evaluated on the original dataset. In contrast, the right graph depicts the rates of CI users and detectors on the CI-simulated dataset.

For CI users, tailored detection tools that focus on subtle auditory inconsistencies could aid their detection. Additionally, incorporating deepfake detection training into CI rehabilitation programs may enhance users' ability to discern manipulated audio. As audio deepfake technologies become more sophisticated, understanding the specific vulnerabilities of different user groups and their differences in perceptual and cognitive strategies is critical for developing inclusive and effective countermeasures.

VI. Humans vs Deepfake Detectors

Recruiting users from marginalized populations can present significant challenges, particularly due to the limited pool of participants willing or able to take part in user studies. To overcome these limitations, CI researchers increasingly rely on CI proxy models [11], [8], [27], [10], [61], [4], [13], which streamline research processes that would otherwise be prohibitively time-consuming and expensive. We seek to assess the practicality of using CI proxies to detect audio deepfakes in security contexts. Specifically, we compare the performance of deepfake detectors using CI-simulated audio with the responses of actual CI users, aiming to determine whether these proxies can serve as cost-effective substitutes for human participants and addressing **RQ3**.

A. Performance Comparison

1) FastAudio and AIR-ASVspoof detectors: We find that the effectiveness of deepfake detection by HPs is similar to the rate achieved by deepfake detectors trained on original audio across real (bonafide), TTS, and VC-generated audio (Figure 11). The discrepancies emerge in the performance between CI users and their respective ML-proxy models. While both (AIR-ASVspoof and FastAudio) deepfake detectors trained on CI-simulated files have near-perfect accuracy on bonafide and TTS-deepfakes, CI users only classified 82% of real audio samples and 67.7% of TTS-generated deepfakes correctly. VC-generated deepfakes posed a much more significant challenge in detection tasks across study participants and ML models, resulting in 35% (AIR-ASVspoof), 56% (FastAudio), and 42.5% (CI users evaluation) correct classification rates.

2) BTS-E detector: While study participants focused on breathing as a cue, they achieved near-perfect detection of real (bonafide) and TTS-generated audio. Similarly, the BTS-E detector, which leverages the inherent complexity of synthetically generating human sounds, such as breathing, when tested against CI-simulated audio, resulted in nearly perfect scores for TTS detection. BTS-E evaluating VC-generated deepfakes also closely resembled that of CI users' FNR.

3) VTR detector: In comparison, the detection via estimation of the vocal tract evaluates deepfake audio independently of the human perceptual cues instead of focusing on the human anatomical feasibility of the vocal tract derived from a given audio sample. By estimating the arrangement of the human vocal tract, the detection focuses on abnormalities that humans cannot perceive directly. Meanwhile, cues specified by participants, such as articulation or pitch changes, would indirectly correlate to the vocal tract estimation, reflecting perceptual anomalies that might coincide with inconsistencies detected by this detector.

B. Model Explainability

Methodology: To investigate the explainability of deepfake detection models, we first decompose audio signals into their fundamental frequencies and amplitudes over time using the Short-Time Fourier Transform (STFT). This method applies a sliding window to separate signals into distinct components. We then transform the data into a Mel-scale spectrogram to align with the non-linear perception of human hearing. Then, we apply saliency maps to identify the regions of the input spectrogram that most influence the model's prediction. Specifically, we use the XRAI algorithm, a modified version of the integrated gradients method [58], to identify the minimal set of spectrogram regions whose removal would significantly decrease the model's confidence in its prediction [32].

To evaluate model behavior, we generate saliency maps for audio samples with and without CI simulation. We analyze ten randomly selected samples for each deepfake attack class, isolating the most influential 30%

Results Figure E1 in Appendix E illustrates the average saliency masks for representative attack types, including TTS (A09) and VC (A17), on both original and CI-simulated audio. For TTS deepfakes, the deepfake detection models primarily rely on features within the 312Hz to 624Hz frequency range (frequency bins 10 to 20) as the most influential for decision-making (Subfigure a&b). These frequency bands correspond to fundamental speech harmonics and energy distributions and are often altered during synthetic speech generation.

Additionally, TTS systems often introduce distortions in these regions, such as imprecise spectral transitions, robotic intonation, flattened dynamics, and irregular formant patterns, along with other simple features. The models exploit these

synthetic artifacts, along with other simplistic features [41], [42], [65], [3], guiding them to easier identification paths. These spectral artifacts are prominent even in CI-simulated datasets, enabling models to maintain reliable detection performance. These same features are detectable by CI users [10], which supports the validity of using models as proxies for TTS deepfake detection when models and users focus on the same auditory regions and patterns.

In contrast, VC systems transform a real human speech sample (source) to mimic another speaker's timbre and vocal characteristics (target) while preserving natural speech dynamics like prosody, pitch variation, and harmonic structures, resulting in audio with fewer spectral distortions. For the original dataset (Subfigure C), the models still focus on the 312Hz to 624Hz range, capturing subtle inconsistencies introduced during VC, such as distortions in pitch alignment, harmonic coherence, or vocal transitions. However, the behavior shifts for CI-simulated VC-generated attacks, as shown in the saliency mask for the A17 attack (Subfigure d). The model no longer prioritizes these upper-frequency bins that are critical in distinguishing VC-generated audio. Instead, it relies on blurred acoustic energy around the bottom three formants, suggesting that the model may be influenced by constant background noise or low-frequency hum in the CI-simulated audio. This shift reduces the model's ability to differentiate between real and deepfake audio, resulting in low model performance on VC-generated deepfakes.

The saliency masks for CI-simulated VC audio highlight a shared limitation between proxy models and CI users: both rely on less informative features. The challenges faced by models on CI-simulated audio closely mirror the struggles observed in CI users during our study. The saliency maps rely on blurred, low-frequency energy rather than nuanced speech patterns, reflecting the perceptual constraints imposed by cochlear implants. When processing VC-generated audio, CI users often depend on prominent auditory cues such as pitch, tone, articulation, and rhythm. However, cochlear implants prioritize lower-frequency information while reducing access to finer spectral and temporal details. These critical features are similarly degraded during the CI simulation process. As a result, both the models analyzing CI-simulated data and CI users shift their focus to less informative features, such as background noise or formant-like artifacts, which significantly diminishes detection performance.

The saliency masks reveal the models' strengths and limitations as proxies for CI users in audio deepfake detection. For TTS deepfakes, the models and CI users align well, relying on easily identifiable artifacts such as robotic intonation and unnatural prosody. Additionally, the CI users' reliance on breathing as a detection cue mirrored the performance of the BTS-E detector, which also leverages the inherent complexity of synthetically generating human sounds, such as breathing.

However, VC-generated deepfakes present a different challenge. Unlike TTS audio, VC deepfakes retain natural prosody and realistic acoustic features, making breathing patterns less effective as a distinguishing cue. This is reflected in the BTS-E detector's performance on VC audio, where its high FNR closely resembles that of CI users, highlighting the shared difficulty in detecting subtle manipulations in VC-generated speech. Saliency maps for CI-simulated VC audio further reinforce this connection, showing that both models and CI users must rely on blurred low-frequency formants and background hum rather than the nuanced speech patterns and breathing cues critical for accurate detection. While this shared reliance reflects some overlap in perceptual limitations, the models fail to fully capture the nuanced strategies and challenges of CI users, especially when subtle auditory cues are critical. These findings highlight the need to better represent fine-grained acoustic details and improve proxy models by incorporating biologically inspired features and saliency-driven insights. By addressing these gaps, ML models could serve as more reliable proxies for CI users, particularly for complex deepfake types like VC, enabling more effective and inclusive detection strategies.

VII. Discussion

A. Human Detection and System Design

Our findings reveal that CI users' abilities to detect audio deepfakes are not uniform across all generation techniques, primarily due to the fundamental differences between TTS and VC methods. This disparity also underscores the broader vulnerability of human perception to various forms of speech synthesis. Specifically, while CI users can reliably detect deepfakes generated by popular TTS techniques, they are significantly more susceptible to VC deepfake attacks [33].

The challenges with VC deepfakes arise from their ability to retain natural qualities from the original audio sample, including prosody, rhythm, emotional depth, and subtle speech characteristics. This preservation makes VC-generated audio more difficult for humans - including CI users- to distinguish from real speech. In contrast, TTS deepfakes often lack these nuanced qualities, producing synthetic artifacts that are more easily detectable.

The widespread familiarity with TTS-generated voices, due to the ubiquity of voice assistants, has likely shaped a subjective bias among listeners. People often associate "deepfake" with the synthetic qualities typical of TTS, such as monotone speech or robotic pacing. While this bias aids in detecting TTS deepfakes, it leaves listeners vulnerable to more sophisticated VC deepfakes where such artifacts are less prominent. As a result, individuals may form internal expectations of how a deepfake "should" sound, leading them to overlook the subtler, more realistic alterations introduced by advanced VC technology [23]. This psychological bias underscores the need to broaden awareness about the evolving capabilities of deepfake technologies.

The discrepancy between human listeners and ML models in detecting deepfakes stems from the different features each relies upon. Humans detect deepfakes based on a subjective interpretation of natural speech, focusing on abnormalities or deviations from expected speech patterns - high-level perceptual features like breathing, rhythm, and articulation.

Consequently, they may become suspicious when audio sounds unnatural based on their prior experiences and expectations. Many study participants relied on these perceptual cues and familiarity to quickly determine audio authenticity. As TTS audio often produces irregularities not found in human speech, participants were able to detect such deepfakes quickly and reliably.

Some deepfake detectors, such as BTS-E, mimic this human approach by analyzing the relationship between breathing, talking, and silence. By focusing on human-detectable features, such models can yield comparable results to humans on these cues. In contrast, other models like AIR-ASVspoof or FastAudio are designed to detect inconsistencies in the frequency domain of the generated audio—subtle anomalies often undetectable to the human ear. While these models can be effective, they may not align with human perceptual strategies. Saliency maps provide a valuable method for evaluating how closely proxy models align with human perceptual strategies. By identifying areas of alignment and divergence, researchers can refine proxy models to better simulate the auditory challenges faced by CI users, particularly for complex deepfake types like VC.

Importantly, many current deepfake detectors are often overfitted to specific datasets, resulting in poor generalization to unseen data. Detection models frequently prioritize achieving a low EER, which can inadvertently increase false positives and cause real audio to be misclassified as fake. An elevated false alarm rate may lead to notification fatigue among detectors' users, diminishing the effectivness of deployed detection systems. For CI users, who already face challenges in identifying natural speech, such an approach could exacerbate difficulties rather than alleviate them.

Developing user-friendly and widely available tools to identify and flag audio deepfakes is critical for mitigating potential threats to CI users and the general population. The currently existing audio deepfake detectors, as suggested by previous research [1] can be easily circumvented. Additionally, some detection methods are complex and require specific technical expertise, limiting their practical utility. The lack of essential resources, such as code repositories or datasets [49], further hinders the widespread adoption of these solutions.

Understanding the specific needs and vulnerabilities of different user groups is crucial for developing effective detection tools and mitigating the most severe risks. By focusing on attacks that can cause disproportionate harm—such as VC attacks targeting CI populations—we emphasize the necessity of recognizing and addressing these threats. Specifically, we recommend concentrating on features that indicate synthetic audio in higher frequency bins, where CI users have a lower detection rate. Gaining insights into how CI users perceive and interact with deepfakes enables us to tailor defense strategies more effectively, moving beyond a one-size-fits-all approach to deepfake detection.

Our user study revealed that CI users classified more audio as real compared to HPs, indicating a higher susceptibility to deception. This finding underscores the necessity of educating particularly susceptible groups about the dangers posed by deepfakes. Raising awareness among CI users and correcting misconceptions about the capabilities of deepfake audio can be an initial protective measure. However, education alone is insufficient to address the broader issue, especially as deepfake technologies continue to advance.

Since conducting human studies is time-consuming and cost-ineffective, using machine learning algorithms to assess human performance offers practical solutions for analyzing large datasets of deepfakes and evaluating effective defensive strategies. Employing proxy models that recognize the most vulnerable aspects affecting susceptible groups would be especially advantageous. These models can facilitate the development of more inclusive and effective detection systems, ensuring that protective measures are attuned to the needs of those most at risk.

Our findings thus emphasize the need for a multifaceted approach to deepfake detection that considers both technological advancements and human perceptual biases. By integrating insights into how CI users perceive and interact with deepfakes, and by developing tools that align with human perceptual strategies, we can create more robust defenses against these emerging threats. This approach ensures that protective measures are effectively tailored to those most vulnerable, addressing the unique challenges faced by CI users and the general population alike.

B. Limitations

Dataset: The ASVspoof2021 audio dataset and its previous iterations are commonly used for deepfake detection evaluation [62], [70]. However, they have limitations in their scope and realism. Most samples are short (i.e., less than 2 seconds) and do not imitate human conversation. The dataset audio quality can also introduce bias within the user study. While this is a good assessment of detection capabilities, it is not necessarily a real-world scenario. It is possible that with longer synthetic speech samples, deepfake audio susceptibility would decrease. Also, this dataset includes more deepfake audio samples than real audio, which may lead to false positive results within real-world applications.

Deepfake Detectors: While we select a variety of synthetic audio detectors for this study, some are excluded due to practical concerns such as unclear or incorrect documentation, or failure to perform as well as expected on the original data. The current selection of synthetic audio detectors provides good coverage of the various types of detectors available.

Priming Bias: Prompting participants to distinguish between deepfakes and authentic human speech may introduce a priming bias by alerting them to the possibility of encountering manipulated content. In real-world scenarios, individuals are typically less aware of being specifically targeted by deepfake audio, which could result in lower detection rates. In contrast, our study explicitly informed participants of the existence of deepfakes, potentially leading to heightened awareness and inflating detection rates by increasing participants' vigilance or skepticism, resulting in an accuracy overestimation.

To mitigate this priming bias, we implemented several measures: first, we avoided showing participants any examples of deepfake content before the task, minimizing their familiarity with specific manipulations. Second, we randomized the order of presentation for both deepfake and authentic audio samples to prevent participants from identifying patterns in their responses. Third, we varied the proportion of each participant's fake and real audio samples, ensuring an unpredictable distribution. Lastly, we emphasized in the instructions that authentic and manipulated audio samples were equally likely, encouraging a neutral evaluation rather than an assumption of deepfake prevalence.

Despite these efforts, the potential impact of priming bias remains a limitation of our study. Future research should consider using more neutral conditions, where participants are unaware of the risk of deepfakes, to better approximate real-world detection capabilities.

CI Simulation: CI simulators are not perfect representations of what CI users hear, nor can they be, as there is significant variation in CI quality and implant techniques. As presented in the study, the ability of a CI user to distinguish deepfake audio speech from natural speech varies widely. Due to the minority status of the CI user population, not as many subjects were found to participate in the survey as desired. Also, a limited number of participants constrained our analysis of deepfake detection differences across CI device types and CI fitting types. While this study engages participants with different CI types, we analyze the average CI user performance and overall group tendencies. Future research could explore a more extensive analysis to examine the effects of implementation to deliver an accurate proxy for CI user detection.

Screening for CIs: The Prolific platform demographic prescreening filters pose limitations due to user self-reporting on each criterion. While we approve and compensate all subjects for a complete submission in Prolific, we validate the Prolific prescreening filter of using a CI in our survey. Such a procedure allows us to remove data from subjects who, at the time of the survey, did not have a CI or potentially misunderstood or answered the platform question unintentionally. However, we rely on subjects' honesty to determine their CI use. Future studies may seek to create a more robust filtering mechanism to establish CI use.

Audio Delivery Method: We do not control for participants' audio delivery methods. Since recruitment occurred via online platforms, participants used their own devices for audio delivery, resulting in variability in the types of headphones or speakers employed. This variability influenced participants' ability to detect deepfake audio by introducing potential inconsistencies in audio playback quality. We did not impose specific requirements regarding headphones or speakers, as the study aimed to reflect real-world conditions where individuals rely on their everyday audio devices. Although the same audio was presented to all participants, the diversity in playback devices introduces an uncontrolled variable that could have affected the results. Future research may consider implementing more controlled experimental conditions, such as providing standardized audio equipment to participants, to mitigate this variability.

A relatively small sample size of CI users can present several statistical analysis limitations, affecting the results' reliability and generalizability. Such a small sample size can increase Type II errors, leading to a high incidence of false positives. Small sample sizes often overestimate effect sizes, producing a wider confidence interval. The relatively low number of CI recipients, the devices' high cost, and the need for individual treatment further contribute to difficulties in gathering large homogeneous samples.

C. Future Work

Audio deepfake technology is constantly evolving, so improved detectors and tools will be needed to prevent attacks. Including state-of-the-art tools in browser extensions and mobile apps could allow deepfake detection to have a practical application for potentially at-risk CI users. Education can also be crucial in making users aware of deepfakes and helping them recognize the potential signs of deceptive content. Improvements in CIs, tuning, and software focusing on improving the perception of speech prosody features would decrease the chances of successful audio deepfake attacks.

VIII. Conclusion

Audio deepfakes have significantly affected our perception and trust in the authenticity of the audio content we encounter daily. Despite advancements in detection methods, current research often overlooks marginalized or underrepresented groups. We investigated the impact of audio deepfakes on CI users and demonstrated their increased susceptibility. Our findings reveal that while TTS deepfakes are generally easier for CI users to detect, VC deepfakes pose a significantly more significant threat.

By integrating ML models with insights into the audio perception of CI users, we underscore the urgent need to enhance deepfake detection models. Our focus on the perspectives of CI users aims to highlight their unique concerns and vulnerabilities, encouraging future research to develop more inclusive and effective detection strategies. Addressing these challenges is crucial for safeguarding both CI users, and the broader population as audio deepfake technologies continue to evolve.

Acknowledgments

This work was supported by the US National Science Foundation under grants CNS-2206950 and CNS-1933208 and the Office of Naval Research under grant N00014-20-1-2205.

References

[1] Z. Almutairi and H. Elgibreen, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," *Algorithms*, 2022.

[2] F. Alvarez, D. Kipping, and W. Nogueira, "A computational model to simulate spectral modulation and speech perception experiments of cochlear implant users," *Frontiers in Neuroinformatics*, vol. 17, p. 934472, 2023.

[3] L. Attorresi, D. Salvi, C. Borrelli, P. Bestagini, and S. Tubaro, "Combining automatic speaker verification and prosody analysis for synthetic speech detection," in *International Conference on Pattern Recognition*, 2022, pp. 247–263.

[4] S. A. Ausili, B. Backus, M. J. Agterberg, A. J. van Opstal, and M. M. van Wanrooij, "Sound localization in real-time vocoded cochlear-implant simulations with normal-hearing listeners," *Trends in hearing*, 2019.

[5] L. Blue, K. Warren, H. Abdullah, C. Gibson, L. Vargas, J. O'Dell, K. Butler, and P. Traynor, "Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction," in *USENIX Security*, 2022.

[6] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, 2006.

[7] T. Brochier, J. Schlittenlacher, I. Roberts, T. Goehring, C. Jiang, D. Vickers, and M. Bance, "From microphone to phoneme: an end-to-end computational neural model for predicting speech perception with cochlear implants," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 11, pp. 3300–3312, 2022.

[8] E. D. Casserly, "Effects of real-time cochlear implant simulation on speech production," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 2791–2800, 2015.

[9] A. Cooper, "Hear me out: Hearing each other for the first time: The implications of cochlear implant activation," *Missouri medicine*, 2019.

[10] H. E. Cullington and F.-G. Zeng, "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects," *The Journal of the Acoustical Society of America*, vol. 123, no. 1, pp. 450–461, 2008.

[11] M. Cychosz, M. B. Winn, and M. J. Goupell, "How to vocode: Using channel vocoders for cochlear-implant research," *The Journal of the Acoustical Society of America*, vol. 155, no. 4, pp. 2407–2437, 2024.

[12] T.-P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, "Bts-e: Audio deepfake detection using breathing-talking-silence encoder," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] M. F. Dorman, P. C. Loizou, A. J. Spahr, and E. Maloff, "A Comparison of the Speech Understanding Provided by Acoustic Models of Fixed-Channel and Channel-Picking Signal Processors for Cochlear Implants," *Journal of Speech, Language, and Hearing Research*, 2002.

[14] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. Gales, and K. Knill, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *IEEE ICASSP*, 2012.

[15] M. D. Fletcher, N. Thini, and S. W. Perry, "Enhanced pitch discrimination for cochlear implant users with a new haptic neuroprosthetic," *Scientific Reports*, 2020.

[16] Forbes, "Fraudsters Cloned Company Director's Voice In $35 Million Heist, Police Find," 2023. [Online]. Available: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=259569357559

[17] J. C. Frank and L. Schönherr, "WaveFake: A Data Set to Facilitate Audio Deepfake Detection," *NeurIPS*, 2021.

[18] Q.-J. Fu, R. V. Shannon, and X. Wang, "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," *JASA*, 1998.

[19] Q. Fu, Z. Teng, J. White, M. E. Powell, and D. C. Schmidt, "Fastaudio: A learnable audio front-end for spoof speech detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3693–3697.

[20] E. Gale, "Exploring perspectives on cochlear implants and language acquisition within the deaf community," *Journal of Deaf Studies and Deaf Education*, 2011.

[21] Y. Gao, R. Singh, and B. Raj, "Voice Impersonation Using Generative Adversarial Networks," in *IEEE ICASSP*, 2018.

[22] J. Gauer, "Audio signal processing methods for the enhancement of music perception in cochlear implant listeners," Ph.D. dissertation, Dissertation, Bochum, Ruhr-Universität Bochum, 2023, 2023.

[23] C. Han, P. Mitra, and S. M. Billah, "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*, 2024, pp. 1–14.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR IEEE*, 2016.

[25] healthdirect, "Cochlear Implant," https://www.healthdirect.gov.au/cochlear-implant, 2020, accessed: 2022-03-01.

[26] M. Hoy, "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants," *Medical Reference Services Quarterly*, 2018.

[27] P. Iverson, C. A. Smith, and B. G. Evans, "Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement and duration," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 3998–4006, 2006.

[28] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," in *Interspeech*, 2019.

[29] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms," in *Interspeech*, 2020.

[30] I. Kaate, J. Salminen, J. M. Santos, S.-G. Jung, H. Almerekhi, and B. J. Jansen, ""There Is something Rotten in Denmark": Investigating the Deepfake persona perceptions and their Implications for human-centered AI," *Computers in Human Behavior: Artificial Humans*, 2024.

[31] C. Kang, "F.C.C. bans a.i.-generated robocalls," Feb 2024, accessed: 2024-02-08. [Online]. Available: https://www.nytimes.com/2024/02/08/technology/fcc-ban-ai-robocalls.html

[32] A. Kapishnikov, T. Bolukbasi, F. Viégas, and M. Terry, "Xrai: Better attributions through regions," in *ICCV IEEE*, 2019.

[33] S. Layton, T. Tucker, D. Olszewski, K. Warren, K. Butler, and P. Traynor, "Sok: The good, the bad, and the unbalanced: Measuring structural limitations of deepfake media datasets," in *USENIX Security Symposium*, 2023.

[34] M. A. Lepori and C. Firestone, "Can you hear me now? Sensitive comparisons of human and machine perception," *Cognitive Science*, 2022.

[35] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive tts training with frame and style reconstruction loss," *IEEE/ACM TASLP*, 2021.

[36] F. Lux, J. Koch, and N. T. Vu, "Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech," in *SLT IEEE*, 2023.

[37] K. T. Mai, S. Bray, T. Davies, and L. D. Griffin, "Warning: Humans cannot reliably detect speech deepfakes," *PLOS ONE*, 2023.

[38] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, 2022.

[39] MED-EL, "What Does a Cochlear Implant Sound Like? Reaching Closest to Natural Hearing With MED-EL," https://blog.medel.pro/, 2021, accessed: 2022-07-09.

[40] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM Computing Surveys*, vol. 54, no. 1, 2021.

[41] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *Interspeech 2022*, 2022.

[42] N. Müller, F. Dieckmann, P. Czempin, R. Canals, K. Böttinger, and J. Williams, "Speech is silver, silence is golden: What do asvspoof-trained models really learn?" *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[43] N. M. Müller, K. Pizzi, and J. Williams, "Human perception of audio deepfakes," in *DDAM 2022*, 2022.

[44] M. Murphy, "Fake Biden Robocall message in New Hampshire alarms election experts," Jan 2024, accessed: 2024-02-07. [Online]. Available: https://www.bloomberg.com/news/articles/2024-01-23/fake-biden-robocall-message-in-new-hampshire-alarms-election-experts?embedded-checkout=true

[45] National Institute on Deafness and Other Communication Disorders, "Cochlear Implants Kernel Description," https://www.nidcd.nih.gov/health/cochlear-implants, 2021, accessed: 2021-11-18.

[46] P. B. Nelson and S.-H. Jin, "Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners," *JASA*, 2004.

[47] S. News, "Deepfake audio of Sir Keir Starmer released on first day of Labour conference," 2023. [Online]. Available: https://news.sky.com/story/labour-faces-political-attack-after-deepfake-audio-is-posted-of-sir-keir-starmer-12980181

[48] S. Nittrouer, E. Tarr, V. Bolster, A. Caldwell-Tarr, A. C. Moberly, and J. H. Lowenstein, "Low-frequency signals support perceptual organization of implant-simulated speech for adults and children," *International journal of audiology*, 2014.

[49] D. Olszewski, A. Lu, C. Stillman, K. Warren, C. Kitroser, A. Pascual, D. Ukirde, K. Butler, and P. Traynor, ""Get in Researchers; We're Measuring Reproducibility": A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences," in *ACM CCS*, 2023.

[50] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research," *Journal of Experimental Social Psychology*, 2017.
[51] J. D. Perezgonzalez, "Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing," *Frontiers in Psychology*, 2015.
[52] Prolific, "Prolific · Quickly find research participants you can trust." https://www.prolific.com/, accessed: 2023-05-20.
[53] C. Rumeau, J. Frere, B. Montaut-Verient, A. Lion, G. Gauchard, and C. Parietti-Winkler, "Quality of life and audiologic performance through the ability to phone of cochlear implant users," *Eur Arch Otorhinolaryngol*, 2015.
[54] M. Saini, "Voice Cloning Using Deep Learning," https://medium.com/the-research-nest/voice-cloning-using-deep-learning-166f1b8d8595, 2020, accessed: 2022-03-01.
[55] M. Schuh and M. L. Bush, "Defining Disparities in Cochlear Implantation through the Social Determinants of Health," *Semin Hear.*, vol. 42, no. 4, Dec. 2021.
[56] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, 1995.
[57] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.
[58] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*, 2017.
[59] Z. Teng, Q. Fu, J. White, M. Powell, and D. C. Schmidt, "GitHub Fastaudio," https://github.com/magnumresearchgroup/Fastaudio, accessed: 2022-09-20.
[60] The Guardian, Feb 2024, accessed: 2024-02-07. [Online]. Available: https://www.theguardian.com/world/2024/feb/05/hong-kong-company-deepfake-video-conference-call-scam
[61] C. S. Throckmorton and L. M. Collins, "The effect of channel interactions on speech recognition in cochlear implant subjects: predictions from an acoustic model," *JASA*, 2002.
[62] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, 2019.
[63] L. van den Heuvel, "Generating enjoyable music for ci users using the wave-u-net model," *Radboud University*, 2021.
[64] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, "Gan-generated faces detection: A survey and new perspectives," *arXiv preprint arXiv:2202.07145*, 2022.
[65] X. Wang and J. Yamagishi, "Investigating active-learning-based training data selection for speech spoofing countermeasure," in *SLT IEEE*, 2023.
[66] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, 2020.
[67] P. Wessel and J. F. Luis, "The GMT/MATLAB Toolbox," *Geochemistry, Geophysics, Geosystems*, 2017.
[68] B. S. Wilson, D. L. Tucci, M. H. Merson, and G. M. O'Donoghue, "Global hearing health care: new findings and perspectives," *The Lancet*, 2017.
[69] World Health Organization, "World Report on Hearing," 2021, https://www.who.int/publications/i/item/9789240020481.
[70] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.
[71] F.-G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, "Cochlear implants: System design, integration, and evaluation," *IEEE Rev. in Biomed. Eng.*, 2008.
[72] Y. Zhang, F. Jiang, and Z. Duan, "GitHub AIR-ASVspoof," https://github.com/yzyouzhang/AIR-ASVspoof, accessed: 2022-09-20.
[73] ——, "One-Class Learning Towards Synthetic Voice Spoofing Detection," *IEEE Signal Processing Letters*, 2021.

APPENDIX

APPENDIX A: ASVSPOOF2019 ATTACKS

This section presents an overview of the attack methods employed in the ASVspoof2019 dataset, a benchmark resource designed to assess the robustness of automatic speaker verification (ASV) systems [66]. As described in Section IV-A, the ASVspoof challenge incorporates various spoofing techniques designed to simulate attacks on ASV systems and evaluate their effectiveness against such threats. Within the Logical Access (LA) scenario, 17 distinct attack methods were implemented, spanning text-to-speech (TTS), voice conversion (VC), and hybrid (TTS&VC) systems. Table A1 provides a breakdown of these attack methods, detailing their system types and the methodologies used.

Table A1: Summary of LA spoofing attack approaches included in ASVspoof2019 dataset, along with their corresponding system types and speech generation approaches.

Attack	System	Approach
A01	TTS	neural waveform model
A02	TTS	vocoder
A03	TTS	vocoder
A04	TTS	waveform concatenation
A05	VC	vocoder
A06	VC	spectral filtering
A07	TTS	vocoder+GAN
A08	TTS	neural waveform
A09	TTS	vocoder
A10	TTS	neural waveform
A11	TTS	griffin lim
A12	TTS	neural waveform
A13	TTS&VC	waveform conc. & filt.
A14	TTS&VC	vocoder
A15	TTS&VC	neural waveform
A16	TTS	waveform concatenation
A17	VC	waveform filtering
A18	VC	vocoder
A19	VC	spectral filtering

APPENDIX B: DEMOGRAPHICS AND INFLUENCES ON DEEPFAKE DETECTION

In this section, we examine how demographic factors influence the ability to detect audio deepfakes among HPs and CI users. Figure B1 illustrates detection accuracy across demographic categories such as gender, age, or education level.

Our analysis indicates significant differences in detection performance based on gender. Among CI users, male participants achieve higher accuracy ($p = 0.01687$), whereas female participants perform better among HPs ($p = 0.04916$). Age is another influential factor showing a statistically significant correlation with detection performance in both groups ($p = 0.04042$). Specifically, we observe a negative correlation, indicating that older participants demonstrate lower accuracy, potentially due to age-related declines in auditory perception.

In addition to demographics, CI-related factors such as device type, duration of CI use, and daily usage patterns influence deepfake detection performance. Participants using bilateral CIs show the highest detection accuracy, whereas

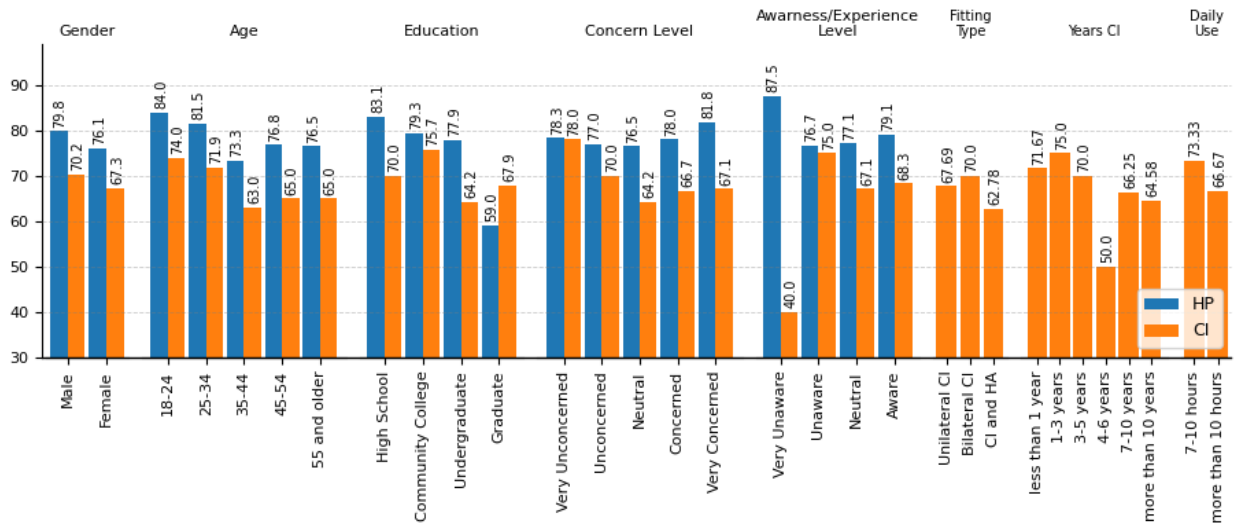


Figure B1: Detection accuracy across different demographic groups for HPs and CI users. Statistically significant correlations are observed between age and performance in both groups. Among CI users, males perform better than females ($p = 0.01687$), while among HPs, females perform better than males ($p = 0.04916$). Additional significant correlations in the CI user group are related to age ($p = 0.04042$) and education level ($p = 0.04185$).

those using CIs in combination with hearing aids (HA) detect deepfakes less effectively. Longer daily CI usage and extended duration since implantation are associated with reduced accuracy. These trends may reflect the combined influence of recipients' age and advancements in implant technologies. These findings provide multifaceted insights into how demographic and device-related factors shape deepfake detection performance. However, the small sample size limited the statistical power to identify significant differences across auditory devices and other CI-related variables ($p < 0.0002$).

Appendix C: Major Emerging Themes Emerging Within Participants' Responses

Table C1: Major categories, key emerging themes, and subthemes from participants' description of audio cues used for classifying audio as real or fake.

Categories	Themes	Subthemes
Speech Prosody Characteristics	Pronunciation & Clarity	Accent, Articulation
	Vocal Expression	Emotion, Tone, Pitch
	Rate & Fluency	Pauses, Rhythm, Speech Rate
Technical Aspects	Sound Quality	Background Noise, Recording Quality
	Vocal Artifacts	Breathing Sounds, Mouth Noises & Nasal Features
Perception & Intuition	Perception	Familiarity, Intuition, Guess
	Intuition	Human, Robotic

We outline the key emerging themes from participants' open-ended responses about the factors they relied on to classify audio samples as real or fake. Following Braun and Clarke [6] thematic analysis framework, as previously detailed in Section V-D, we categorized these responses into major categories, themes, and subthemes. Table C1 presents a structured summary of the findings, offering insight into the perceptual cues participants used in their deepfake evaluations. The diversity of auditory cues identified underscores the complexity of human audio deepfake detection and the challenges in distinguishing real audio from deepfakes.

Appendix D: User Study Protocol

Figure D1 shows the web interface used in the study. Participants were required to listen to a series of 20 audio samples and determine whether each sample was a real human voice (Real) or computer-generated deepfake (Fake), and rate their confidence level on a 5-point Likert scale. They also were asked to provide open-ended responses describing any cues or factors that influenced their judgment.

This study aimed to assess participants' ability to detect audio deepfakes while collecting qualitative insights into the auditory cues they relied on. A standardized protocol ensured all participants had a consistent experience, allowing for systematic analysis of their responses.

Appendix E: Saliency Masks

In this section, we present the saliency maps derived from the original and CI-simulated audio samples from the ASVspoof2019 dataset for TTS- and VC-generated deepfakes. These maps help visualize the most influential regions contributing to the model's classification decisions, providing

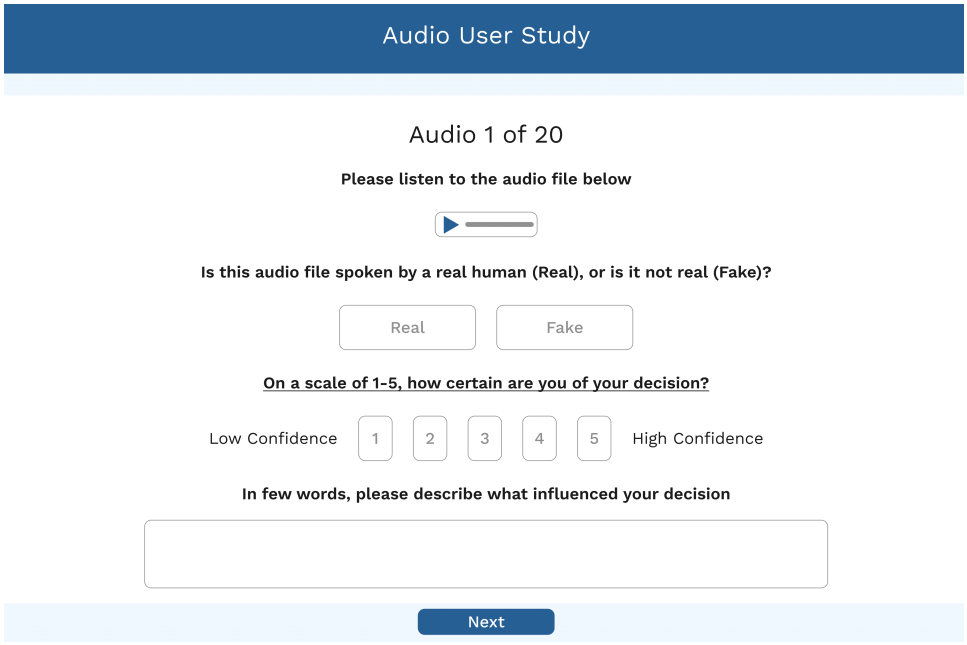
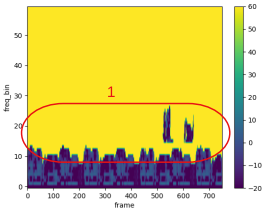
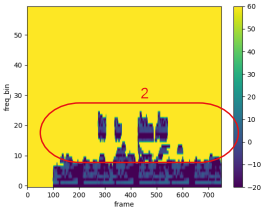


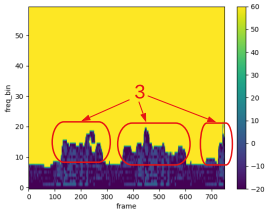
Figure D1: The study interface as given to participants. They were asked to listen to an audio file, decide whether the audio is real or fake, rate their confidence on a 5-point Liker scale, and describe what factors influenced their decision.



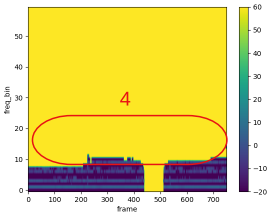
(a) A09 - TTS without CI simulation



(b) A09 - TTS with CI simulation



(c) A17 - VC without CI simulation



(d) A17 -VC with CI simulation

Figure E1: Mean saliency masks for original and CI-simulated audio samples using attacks A09 (TTS) and A17 (VC) with the most salient 30% area.

insights into how the model differentiates spoofed audio from real audio.

Figure E1 shows the mean saliency maps for TTS attack A09 and VC attack A17, both with and without CI simulation. We plotted ten random audio samples for each attack method on Mel spectrograms. Then, we applied the eXplainable AI (XRAI) saliency method [32] to identify the mean top 30% most influential regions for the model's classification.

For the TTS attack (A09), the saliency maps in Figures E1a and E1b indicate that the key prediction areas, particularly artifacts above frequency bin 10 (areas 1 and 2), remain highly silent, even after applying CI simulation on audio, suggesting that the CI simulation has a small effect on the model's ability to detect TTS-generated deepfakes.

In contrast, the saliency maps for the VC attack (A17) (Figures E1c and E1d) show that higher frequency bins lose importance after CI simulation. Specifically, areas above frequency bin 10 (area 3 in the original audio) lose prominence in the CI-simulated version (area 4), indicating that CI simulation covers features the model relies on for detecting VC attacks.