"Better Be Computer or I'm Dumb": A Large-Scale Evaluation of Humans as Audio Deepfake Detectors*

Kevin Warren
University of Florida
Gainesville, FL, USA
kwarren9413@ufl.edu

Tyler Tucker
University of Florida
Gainesville, FL, USA
tylertucker1@ufl.edu

Anna Crowder
University of Florida
Gainesville, FL, USA
annacrowder@ufl.edu

Daniel Olszewski
University of Florida
Gainesville, FL, USA
dolszewski@ufl.edu

Allison Lu
University of Florida
Gainesville, FL, USA
allison.lu@ufl.edu

Caroline Fedele
University of Florida
Gainesville, FL, USA
cfedele@ufl.edu

Magdalena Pasternak
University of Florida
Gainesville, FL, USA
mpasternak@ufl.edu

Seth Layton
University of Florida
Gainesville, FL, USA
sethlayton@ufl.edu

Kevin Butler
University of Florida
Gainesville, FL, USA
butler@ufl.edu

Carrie Gates
Dalhousie University
Halifax, Nova Scotia, Canada
cegates@dal.ca

Patrick Traynor
University of Florida
Gainesville, FL, USA
traynor@ufl.edu

ABSTRACT

Audio deepfakes represent a rising threat to trust in our daily communications. In response to this, the research community has developed a wide array of detection techniques aimed at preventing such attacks from deceiving users. Unfortunately, the creation of these defenses has generally overlooked the most important element of the system - the user themselves. As such, it is not clear whether current mechanisms augment, hinder, or simply contradict human classification of deepfakes. In this paper, we perform the first large-scale user study on deepfake detection. We recruit over 1,200 users and present them with samples from the three most widely-cited deepfake datasets. We then quantitatively compare performance and qualitatively conduct thematic analysis to motivate and understand the reasoning behind user decisions and differences from machine classifications. Our results show that users correctly classify human audio at significantly higher rates than machine learning models, and rely on linguistic features and intuition when performing classification. However, users are also regularly misled by pre-conceptions about the capabilities of generated audio (e.g., that accents and background sounds are indicative of humans). Finally, machine learning models suffer from significantly higher false positive rates, and experience false negatives that humans correctly classify when issues of quality or robotic characteristics are reported. By analyzing user behavior across multiple deepfake datasets, our study demonstrates the need to more tightly compare user and machine learning performance, and to target the latter towards areas where humans are less likely to successfully identify threats.

CCS CONCEPTS

• **Security and privacy** → *Usability in security and privacy*; • **General and reference** → **Evaluation**; **Measurement**; • **Computing methodologies** → Cross-validation.

KEYWORDS

audio deepfake, user study, security, classification

ACM Reference Format:
Kevin Warren, Tyler Tucker, Anna Crowder, Daniel Olszewski, Allison Lu, Caroline Fedele, Magdalena Pasternak, Seth Layton, Kevin Butler, Carrie Gates, and Patrick Traynor. 2024. "Better Be Computer or I'm Dumb": A Large-Scale Evaluation of Humans as Audio Deepfake Detectors. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3670325

1 INTRODUCTION

Audio deepfakes allow nearly anyone to create human-sounding speech without the actual existence or consent of a real person. While such audio can have many beneficial uses [58], the potential to use deepfakes in service of fraud [20, 25] or disinformation [59, 67] is significant. Given the early success of such efforts, it is highly

*The title comes from the free response portion of our study.

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

ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3670325

likely that users will be increasingly exposed to deepfake audio in the business, social, and political parts of their lives.

As a means of spurring innovative and strong defenses against these threats, the research community has developed multiple public datasets for testing. Composed of a wide array of transformed and entirely synthesized voices, these samples are designed to provide community benchmarks. However, as is common in the security community, a crucial component of detection has been left out of the design - users. As such, because a large-scale study of human performance against such samples has not taken place, it is not clear to what extent the current state-of-the-art samples used to build detectors actually fool humans. Moreover, the extent to which detectors correctly classify attacks missed by users is also not understood. Without this grounding in user performance, it is unclear how currently proposed defenses can be meaningfully incorporated into real-world systems.

We address the above issues by performing the largest multi-dataset study of audio deepfakes. Our efforts produce the following contributions:

- **Largest user study on audio deepfake detection:** We conduct a user study with over 1,200 participants using statistically parameterized, stratified sampling from the three most widely-cited audio deepfake datasets - Wavefake [21], ASVspoof2021 [77], and FakeAVCeleb [35]. Across these datasets, we show that humans exhibit a 73% accuracy, with an elevated ability to correctly identify other humans over deepfake audio.
- **Qualitative study identifying decision factors:** As part of our study, we asked users to explain their classification decisions. We collect and manually categorize over 24,000 responses into themes, thereby providing insight into how humans classify audio as fake or real.
- **Comparative analysis on human and ML performance:** Without a clear understanding of why some samples fool users better than others, it is challenging to design defenses that actually help them. Having characterized the dataset from the user perspective, we identify differences in how models perform on the same inputs, and identify possible areas of improvement for future efforts. We also make our annotated dataset publicly available such that future researchers can see how their mechanisms perform against user classification without the significant expense of a user study.

We aim to characterize how well people perform on discriminating deepfake datasets and how this compares to machine learning detection models. Toward this goal, we ask the following research questions:

RQ1 **Evaluate Human Accuracy:** What are the performance metrics for humans when discriminating on the popular datasets used in audio deepfake detection?

RQ2 **Determine Human Classification Reasonings:** What are the common themes affecting how humans classify audio deepfake samples as real or fake?

RQ3 **Compare Human and ML Classification:** Is there a demonstrable difference in audio deepfake detection capability between humans and ML models? Furthermore, are there common themes amongst these differences?

The remainder of this paper is organized as follows: Section 2 covers background material and related work; Section 3 discusses the datasets used and our methodology for evaluating them; Section 4 presents the quantitative results of our user study; Section 5 provides qualitative and thematic results from our user study; Section 6 compares user performance to four benchmark ML detectors; Section 7 provides discussion and recommendations; Section 8 offers concluding remarks.

2 BACKGROUND AND RELATED WORK

Deepfake Detection vs. Speaker Verification: There is a large overlap between deepfake detection and speaker verification tasks; however, they are not the same. It is a fundamentally different task to verify the generative source of audio than to verify the identity behind the voice in an audio clip. When a deepfake is targeted to impersonate a specific individual, the recipient may use prior knowledge of that individual's voice to help determine the validity of the audio they are hearing. It is in this specific case that these two tasks become largely the same to the end recipient.

While some deepfakes are targeted at familiar voices and largely fall under this speaker verification task, there are instances where the recipient is not familiar with the voice of the deepfake but still needs to determine the validity of the voice. This is often seen in cases such as call centers, where representatives typically have no familiarity with the person on the other side of the phone call [53]. In these cases, the representative is tasked with a pure deepfake detection problem and needs to appropriately determine if the person on the line is human or computer-generated for the security of their customers. Without prior knowledge of the person's voice, call centers must rely on other artifacts to determine if the voice is a deepfake. While the distinction between these two tasks can be minute, it is important to understand the differences to appropriately categorize results and identify limitations and/or restrictions available to the people doing the classification.

Audio Deepfake Detection: Research on audio deepfake detection identified many characteristics of human audio that differentiate it from synthetic audio. These characteristics include the airflow pressure or time-difference-of-arrival of phoneme sequences [73, 81], the presence of breathing [37], the pop sound made by a breath [69], the particular attributes of the airwaves [7, 68], the movement or structure of the human vocal anatomy [8, 80], and even subtle spectral differences [4, 43]. Some of these techniques require additional hardware to implement [68, 73] while others can be implemented using the hardware that already exists on a mobile device [69, 80]. Many researchers and organizations have synthesized audio deepfake datasets [21, 35, 41, 50, 63, 77] to advance these detection techniques, especially those relying on machine learning models [10, 70, 79]. Additionally, recent work suggests that these datasets can improve by considering the base rate of deepfakes in the wild [38]. Although deepfake detection systems continue to make advancements, significantly less research has been done on how humans detect deepfakes.

Studies Outside of the Audio Domain: Several studies into the capabilities of humans to detect generated media have been performed outside of the audio domain. These studies investigated how people evaluated and interacted with deepfake videos, deepfake

images and generated text. Studies performed for video deepfake detection [22, 36] aimed to compare human accuracy with the leading detection models [11, 19, 48] to determine if there was a difference in the capabilites of humans and machines to identify deepfakes. The studies performed within Natural Language Generation were focused on addressing the variability in how machine generated text was evaluated [66] and determine the efficacy of potential detection mechanisms [16] compared to human evaluation. Mink et al. [47] combined machine generated content from two mediums, text and images, when they investigated how users observe fake profiles on social media.

Studies on Audio Deepfakes: Unlike other generated media, there have been limited studies into the way humans perceive audio deepfakes. The majority of these studies are performed within the context of voice impersonation attacks [49, 52, 74], where the audio generated are voice conversion techniques [29, 56] performed on either celebrities or participants from another user study. These studies tested participants with a speaker verification problem by giving them side-by-side comparisons of voices they should be familiar with and tested them against speaker authentication systems [1, 46]. Wenger et al. [74], contextualized the problem by demonstrating that people are more susceptible in a work setting compared to personal time. Additionally, research suggests that the lack of differentiable brain activity when processing real and synthetic audio suggests that humans will inevitably lose the ability to reliably classify synthetic audio once it reaches a high level of quality [52].

These studies either create their own set of deepfakes [52, 74] or perform their study on only a single dataset [42, 51]. Both of these options limit the applicability of the results and *do not fully capture the capabilities of audio deepfakes to impact people*. While these studies are a good initial starting point for studying audio deepfakes, they are not expansive or comprehensive enough to make meaningful recommendations or draw conclusions. Comparatively, our study tests human classification across the three most popular and cited audio deepfake datasets, which represent the state-of-the-art for audio deepfakes in the research community. We perform our study using a principled, statistical approach to appropriately evaluate human detection capabilities on each dataset.

3 METHODOLOGY

To explore our research questions, we conduct an online user study that tests the capacity of humans to act as audio deepfake detectors. We use Prolific over MTurk due to MTurk responses lacking generalizability compared to Prolific [65]. For our experiment, we test participant detection capabilities against a subset of samples from the three datasets described in Section 3.1.

We design a statistically principled user study to use human classification to contextualize datasets within real-world scenarios based on the following null hypothesis:

H_0 · humans will classify deepfake audio files at the same rate that they classify human audio.

In the following sections, we describe our dataset selection decisions, audio sampling methodology, experimental design, ethical considerations, and participant recruitment.

3.1 Datasets

For our study, we collect samples from the three most popular audio deepfake datasets: ASVspoof2021, Wavefake, and FakeAVCeleb. These datasets together exhibit a high citation count paired with a variety of generation techniques, sample durations, and speaker counts. Consequently, we believe these three datasets together act as a proxy for the corpus of English samples within the audio deepfake research area.

Wavefake: The Wavefake dataset represents a single speaker saying the same sentences in both the real and deepfake samples. It is a deepfake dataset that contains ten sets of deepfakes using six different generation architectures across two different languages. The dataset is primarily developed around the LJSPEECH corpus [28] for its English samples and the JSUT speech corpus for the Japanese deepfakes [63]. For the purposes of our study, we only sample from the English portion of the dataset. The LJSPEECH dataset contains 13,100 short audio clips read by a single female English speaker. Since its release in 2021, Wavefake has been used in the development of several detection algorithms [31–33, 72] and evaluation of existing algorithms [64, 71].

ASVspoof2021: The deepfake dataset from the ASVspoof2021 competition [77] contains samples that represent multiple unfamiliar speakers saying a variety of phrases, and is split into three tracks: physical access data, logical access data, and deepfake data. For the purposes of our study, we focus only on the deepfake track which uses text-to-speech (TTS) and voice conversion (VC) generation methods. We use the evaluation dataset from the deepfake track, which contains 22,617 human samples and 589,212 deepfake audio samples that are generated using more than 100 different audio spoofing algorithms and processed using various lossy codecs. ASVspoof2021 has enabled numerous works [5, 40, 75] to advance the field of audio deepfakes since its inception.

FakeAVCeleb: FakeAVCeleb [35] is an audio-video deepfake dataset that contains deepfake videos as well as lip-synced fake audio. It represents a variety of familiar voices with samples from celebrities saying a multitude of phrases. The dataset was formed using YouTube videos of celebrities from four different ethnic backgrounds sampled from the VoxCeleb2 corpus [13], a dataset containing over one million YouTube videos of 6,112 celebrities. To produce deepfake audio, the developers use the voice cloning tool SV2TTS [30] as their sole generation algorithm. For our study, we collect the 10,209 real audio samples and 11,335 deepfake audio samples used within the dataset. Numerous studies have used FakeAVCeleb in the audio domain to test existing models [55] or train new models [31, 32]. Others have used it in the combined audio and visual domains to perform similar tasks [9, 12, 14, 17, 18, 23, 26, 27, 34, 39, 61, 76, 78].

3.2 Audio Sampling

Unlike evaluating an ML classifier on a large dataset, scaling is cost-prohibitive for human subjects. To narrow down the task, we calculate confidence intervals [3, p. 362-365] to estimate the total number of samples in each dataset that we need to discriminate on to achieve a population accuracy within a desired error range. Under the assumption that each dataset is a large population of samples (i.e., total samples » 1000), we can formally estimate our

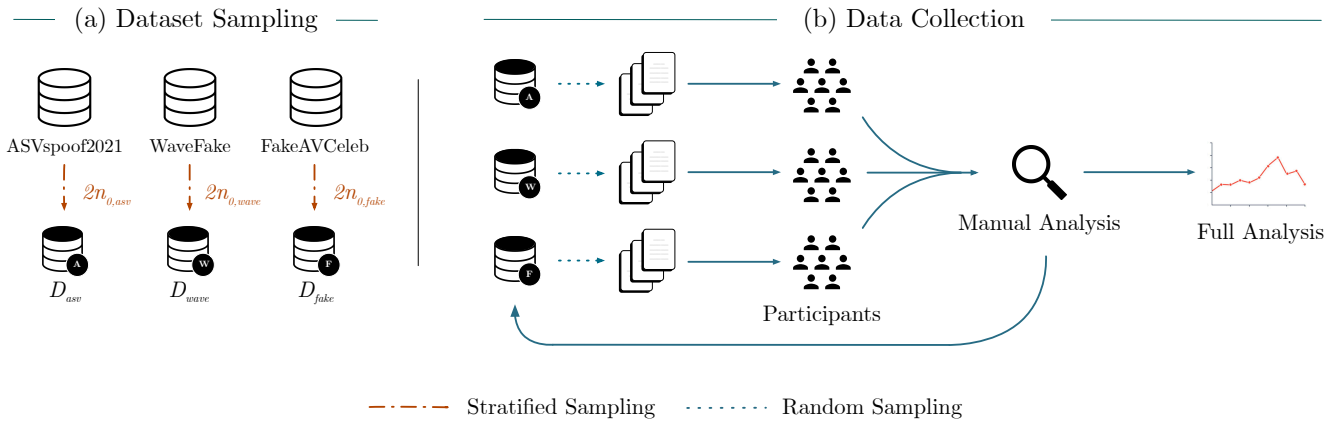

Figure 1: We begin construction of our survey by performing stratified sampling on our three datasets to extract $2 * n_0$ samples per dataset to create $D_{\{wave,asv,fake\}}$. Then, for each participant, we randomly select 20 samples from one of these populations. Each sample is listened to by at least three unique participants. We discard bad responses and provide those samples to a new user until we receive three valid responses for every sample. We conclude our survey and store our results in a database for analysis.

sample count n:

$$n = \frac{z^2 \cdot q(1-q)}{ME^2}, \tag{1}$$

with q as the estimated study accuracy, ME as the desired margin of error, and z as the z-score. Under the assumption that the dataset is balanced between the classes (i.e., there are approximately the same number of real and deepfake samples), Eq. 1 provides a close estimate for sample count. In reality, however, these datasets are not always balanced and generally favor the deepfake class (e.g., 97% deepfake samples to 3% real samples for ASVspoof2021). To account for the imbalance, we take the original margin of error equation

$$ME = z_{\alpha/2}\sqrt{Var(q)}, \quad Var(q) = \frac{q(1-q)}{n}, \tag{2}$$

used in Eq. 1, and adjust the variance to account for the different class proportions as follows:

$$Var(q) = p_{DF}^2[\frac{q_{DF}(1-q_{DF})}{n_0}] + p_H^2[\frac{q_H(1-q_H)}{n_0}], \tag{3}$$

where q_{DF} is the study accuracy on deepfakes, q_H is the study accuracy on real human audio, p_{DF} is the proportion of the dataset that is deepfakes, p_H is the proportion of the dataset that is real human audio, and n_0 is our subclass sample size.

To simplify the sample size calculations, we assume that the human participants classify deepfakes and human audio at the same rate (i.e., our null hypothesis). We can solve the margin of error equation with the new variance for n_0 to get

$$n_0 = \frac{z_{\alpha/2}^2}{ME^2}(p_{DF}^2 * q_{DF}(1-q_{DF}))(p_H^2 * q_H(1-q_H)). \tag{4}$$

We assume a 95% confidence interval (i.e., $z_{\alpha/2} = 1.96$), a desired margin of error of 2%, and an estimated q_{DF} and q_H of 0.75 based on previous similar work [51]. Solving for n_0 only gives us the number of samples needed for the survey for one of the subclasses, either human audio or deepfakes. We want the participants to have an equal chance of receiving a human or deepfake sample on each question so as not to bias the classification, thus we use stratified sampling to match the number of deepfakes with real human audio to get a total sample number of $2n_0$ for each dataset. The data distribution for the datasets are as followed: **Wavefake** – 89% deepfake / 11% real; **ASVspoof2021** – 97% deepfake / 3% real; **FakeAVCeleb** – 53% deepfake / 47% real. Solving Eq. 4 for each of the distributions, we determine that the total number of samples needed for each dataset are 3,400 for ASVspoof2021, 2,880 for Wavefake, and 1,800 for FakeAVCeleb. We refer to the sampled audio as $D_{\{wave,asv,fake\}}$, where the subscript denotes the dataset.

3.3 Experimental Design

Our user study evaluates human classification of audio deepfakes. The only control we impose on the design is the separation of the three datasets (i.e., participants only receive samples from a single dataset). Each participant is instructed to listen to twenty audio samples on separate pages, and answer a set of questions. Each audio sample is pulled from a pool of unique files for a single dataset. We share an overview of our experimental design in Figure 1.

Study Procedure: Each Prolific user who signs up for our survey is redirected by a link to our survey website. They start by reading a description of the survey, a consent form, and a General Data Protection Regulation (GDPR) addendum, and are asked to continue if they understand and give consent. Before participants begin the study, we confirm that they do not have any hearing difficulties or loss and ask them to verify they are using headphones for the survey (a requirement described in the survey description). We ask participants to use headphones to reduce both noise and distractions as well as to standardize the experience of each person taking the survey.

Then the participants are asked to do the main task of discriminating audio samples. For each of the twenty samples, they are given an audio clip to play accompanied by three questions:

(1) Was the sample Human or Computer generated?
(2) What is your confidence level with your decision? (1-5)
(3) Did you hear anything that affected your decision?

The term "deepfake" carries a certain stigma with it that can bias the way that people think or interact with media [59]. Consequently, we avoid using this term anywhere throughout the survey and explicitly only use the term "Computer generated" instead. We also give the participants two attention checks to verify that they are fully participating in the survey.

After they complete the main task, we ask the participants some follow-up demographic questions and if their concern level with deepfakes changed due to their experience with the dataset.

Survey Sample Selection: We randomly split each dataset into clusters of 20 audio samples and each participant is given a single cluster. We run through each dataset three times and randomly sample new clusters for each run of the survey. This means that each audio sample is classified independently of the other audio and order of the cluster. For example, audio sample A given to participant P1 could be the 5^{th} sample shown, while sample A if given to participant P2 could be the 11^{th} one shown. Additionally, we do not control the distribution of samples each participant receives since the sampling of the audio files from the pool is not stratified. This is done to remove the bias of participants feeling the need to even out their classification decisions. Thus, our data follows a binomial distribution of human and deepfake samples and has a small likelihood that a participant receives all samples from a single category (i.e., human or deepfake).

Survey Response: We want to determine and evaluate how the average person would perform on each dataset. To minimize any outliers in skill level, we recruit enough participants such that each audio sample is heard and evaluated three times. Achieving three participants per audio file allows us to implement a voting scheme, where we assign a decision for each audio sample based on a majority vote. There are several cases in which responses are excluded and replaced (refer to Section 3.4), thus we guarantee three independent responses for each audio file. By using a majority vote, we get an estimation of the performance of an average person and look at the population as a whole.

Self-Training Control: Two control elements help minimize the opportunity participants have to "train themselves" throughout the study: isolation of datasets and lack of feedback. By only giving the participants samples from a single dataset, we limit the exposure they have to other deepfakes which can influence future decisions by comparison. Our study assesses each dataset individually and focuses on the unique experience each dataset presents. We also *do not* inform the participants at any point if their decision was correct. This limits the impact of question ordering affecting later accuracy in this scenario.

Attention Check and Action Verification: To verify the reliability of the results, we implement two attention checks throughout the survey and make sure the required actions are completed. Each attention check requires participants to listen to an audio clip and pick out the appropriate transcription from a multiple-choice list. These checks provide each participant with a trivial task to ensure that the participant is not just randomly picking answers. We also record the number of times each participant listens to each audio sample. Finally, we autoplay the audio on each page and restrict continuation until it has fully completed to ensure each participant hears the entire sample.

3.4 Ethics and Participation

Before recruiting for our survey, we gained approval from our Institutional Review Board (IRB) as an exempt study. Our work was exempt because we collect no personally identifiable information (PII) as Prolific maintains the anonymity of its users. Additionally, any behavior requested from the subjects was in the form of benign written responses. Each participant is shown an informed consent page at the beginning of the survey detailing the task they are performing, the time requirement, the security of their data, and the ability to withdraw. With our institutional exempt approval, there is no requirement for a documented signature, thus we ask participants in the survey to start only if they have read and agreed to the information shown on the consent page.

We use the following three requirements for selecting participants: users without any hearing loss or hearing difficulties, users with English as their first language, and users located in the United States.

We recruited 1,212 participants based on the number of samples required from our sampling calculations detailed in Section 3.2. We checked that the participants met our three requirements and assigned 510 for ASVspoof2021, 432 for Wavefake, and 270 for FakeAVCeleb as this sufficiently covers each audio sample three times based on our calculations in Section 3.2. The median time for completion for all participants was 14 minutes and each participant received $5 in compensation regardless of if we excluded their data. After manual inspection of their responses, we excluded data if they failed an attention check, seemed to give automated responses (e.g., same responses for every question), completed the survey too quickly (< 4 minutes), or reported that they had issues hearing the audio. For each set of excluded data, we re-ran that set of audio files for a new participant until we successfully completed all 1,212 individual surveys. Overall, the cost of compensating participants, hosting our website, storing data, and recruiting additional participants to compensate for poor responses exceeded $10,000. We had participants in every age range from 18 to 55+, with the median age group being 35-44 years old. Approximately 53% identified as male, 44% as female, 2% as non-binary, and 1% identified as either "other" or preferred not to answer. We provide a full list of our demographics on our companion website[1].

4 USER STUDY EVALUATION

In this section, we present the results of our user study including individual human performance on the datasets, human performance based on consensus voting, and model metrics from the voting performance to address **RQ1**.

[1] https://sites.google.com/view/better-be-computer/home

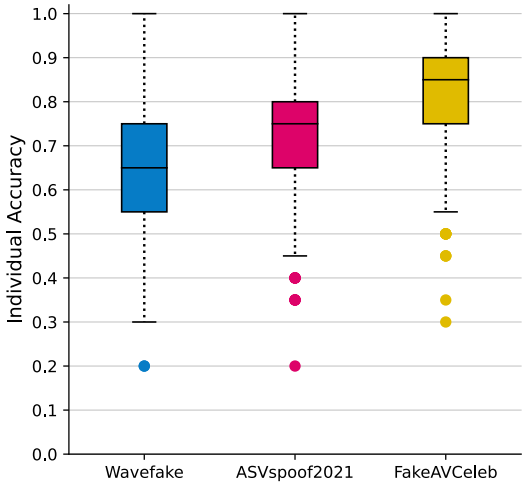

Figure 2: Individual user accuracy on the 20 samples given to each participant. Each dataset had at least one person score a perfect accuracy, however, the average performance varied from dataset to dataset. On average, participants performed better on FakeAVCeleb and worse on the Wavefake dataset.

4.1 Individual User Performance

We first look at the individual performances of the survey participants by looking at their accuracies across the 20 samples given to them. While previous work has looked at this type of performance, they limited their results to a single dataset (e.g., a community dataset or self-created dataset) [42, 51]. Our study provides an expanded perspective on this previous work, looking at three of the most widely used community datasets for audio deepfake detection. A summary of the individual accuracies for each dataset is displayed in Figure 2.

While at least one individual achieved 100% accuracy on each of the datasets, the average performance across the datasets varied. Users performed the worst on Wavefake with a mean accuracy of 65%. Overall performance on ASVspoof2021 was slightly better with a mean of 71%. The highest performance was on the FakeAVCeleb dataset with an average accuracy of 81% and a minimum of only 30% compared to the 20% minimum on the other datasets. The variation in performance between each of the datasets reveals that not all deepfake audio datasets are created equal. The way people process various types of deepfakes differs, so the composition of the dataset has an impact on how susceptible people are to misclassifications.

We additionally investigate if our study provides training to users by plotting individual user accuracy by question number, shown in Figure 3. This plot does not exhibit any demonstrable change in accuracy as users progress through our survey, suggesting that we mitigate training as desired in our experimental design.

4.2 Application of Voting Scheme

Going beyond individual performances, we also assess how the average person would perform across the entirety of each dataset. To accomplish this, our survey collects three responses from different participants for each audio sample in our subsets. Using the three responses, we apply a voting scheme to the classifications to identify how the average person would classify each sample. With only a limited subset of audio, the voting scheme reduces the impact of outlier performance and gives us a more general view of human performance on the samples.

The voting scheme yields two types of decisions: complete agreement and split decisions (i.e., 2 of 3 agreed). The percent of decisions that were complete agreements follows the same trend as the individual user performances; FakeAVCeleb is the highest at 61% complete agreement, followed by ASVspoof2021 at 48%, and the lowest is Wavefake at 39%. Additionally, we observe that the accuracy of complete agreement was approximately a 35% relative increase from the accuracy of split decisions. This demonstrates that when samples seem clearly fake or real to people, people are generally correct (i.e., there are not many fake samples that sound perfectly human and there are not many real sounds that sound demonstrably fake).

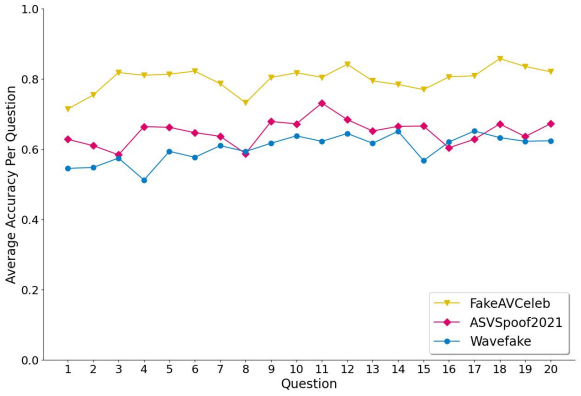

Figure 3: Plot of individual user accuracy throughout our 20 question survey. We observe no correlation between question number and accuracy, which suggests that the presentation of our survey does not train participants.

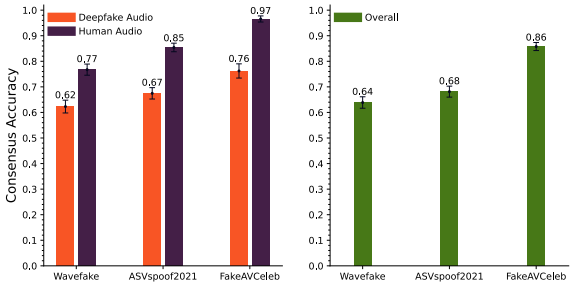

Figure 4: User accuracy based on consensus voting by dataset on human audio versus deepfake audio and human accuracy on the audio overall. Users as a group performed better on human audio across all datasets, and performance on FakeAVCeleb was the highest overall.

	z-score	p-value
Wavefake	8.42	3.78e-17
ASVspoof2021	12.33	6.57e-35
FakeAVCeleb	12.58	2.76e-36

Table 1: Results of the z-test comparing the accuracy of user performance on real audio versus deepfake audio (H_0 : $q_{DF} = q_H$).

4.3 Model Metrics on Human Performance

Using the voting scheme classifications, we can calculate the accuracy of a general person on the subset of samples for each dataset and extrapolate those results using confidence intervals as described in Section 3.2. Figure 4 shows the accuracy performance with the voting scheme applied on both the real and deepfake audio separately as well as their overall performance. Notably, our findings reveal that users consistently exhibit higher accuracy on real audio as opposed to deepfake audio (e.g., 15% better for Wavefake, 18% better for ASVspoof2021, 21% better for FakeAVCeleb). We investigate this difference further by performing a z-test using our original null hypothesis ($H_0 : q_{DF} = q_H$) [3, p. 475-476]. The results of the z-test lead us to reject the null hypothesis for all datasets as each corresponding p-value is below 0.005[2] as shown in Table 1. The results of the z-test are further confirmed by the lack of overlap between the confidence intervals around the real and deepfake audio accuracies as shown in Figure 4.

In computing the overall accuracy, we weight the real audio accuracy and deepfake audio accuracy based on the distribution of human and deepfake audio instances within each dataset. We employ Eq. 3 to quantify the margin of error associated with the overall accuracy. This equation requires that the variables representing human audio accuracy and deepfake audio accuracy are statistically independent. We assume independence when calculating the sample size for our survey, recognizing that any departure from this assumption would result in a margin of error larger than anticipated.

To validate the assumption of independence, we conduct both Spearman's rank and Pearson's correlation tests on the two variables. The highest absolute correlation value was the Spearman's rank ($r = -0.11, p = 0.015$) for ASVspoof2021, suggesting the two variables are dissimilar. Given the observed negligible correlation between human audio accuracy and deepfake audio accuracy across all analyzed datasets, we conclude that these two variables are indeed statistically independent.

We design our study to calculate an overall accuracy that could directly compare with detection models. For comparison with detectors, we treat the human responses as a "human model" and provide a full set of model performance metrics for each dataset in Table 2. Assessing the humans as their own model demonstrates that people tend to lean towards believing a piece of audio is human, with FNR scores consistently higher than FPR scores across all datasets. Additionally, we observe that *all* standard classification

[2]Classical standards for p are < 0.05 while modern standards are p < 0.005. We adopt the latter [6].

	Human "Model" Performance		
	Wavefake	**ASVspoof2021**	**FakeAVCeleb**
Accuracy	63.9% ± 2.2%	68.1% ± 2.4%	85.8% ± 1.6%
Precision	.73	.82	.96
Recall	.62	.67	.76
F_1-Score	.67	.74	.85
FPR	23%	15%	3%
FNR	38%	33%	24%

Table 2: Model performance metrics on the survey responses for each dataset. The highlighted false positive rate (FPR) and false negative rate (FNR) demonstrates that people trend towards trusting audio as being human generated since the FNR is consistently greater than the FPR.

performance metrics follow the same dataset trend as shown in Figures 2 and 4: participant performance is the worst for Wavefake and the best for FakeAVCeleb (e.g., FPR is nearly eight times as high for Wavefake when compared to FakeAVCeleb). These quantitative results address how well humans classify audio from popular deepfake datasets (**RQ1**) and motivate further analysis into why humans conclude that audio is trustworthy or not.

5 THEMATIC ANALYSIS

We perform a qualitative thematic analysis by developing a codebook and coding the text response of our participants to gain further insight into **RQ2**. We describe our codebook-generating process, explain the themes that emerge from our coding, and examine the qualitative breakdown that response reasoning had on performance.

5.1 Coding Process

We perform a thematic analysis of the users' open-text responses to the question "Did you hear anything that affects your decision?" to characterize how humans approach classification of deepfake audio. A codebook is developed by a group of raters through a discussion of common ideas observed during an initial pass of the responses. Keywords from the first pass are grouped together by likeness and eight unique codes are given to represent each group. Note that the codes reflect the general idea behind the group of keywords and do not necessarily follow strict definitions. All codes along with their associated keywords and descriptions are shown in Table 3. Two raters independently code all responses using the eight codes and inter-rater agreement is measured via the Cohen's Kappa coefficient. Each rater also had the option to mark a response as a "Bad Response" or "No Reason" if they believed the response did not appropriately answer the question or indicated a lack of reasoning. We removed all responses in which at least one rater indicated either a "Bad Response" or "No Reason". Finally, if responses included multiple clauses that represent different codes, we only consider the first clause as it represents the participant's initial reaction.

When strong agreement ($\kappa >= 0.8$) [45] is not initially reached, a third rater re-codes a portion of the responses. Responses for re-coding are selected based on a count of codes where the two original raters differ. Because Rater 1 uses the "Human-Like" and "Robotic" categories at a much higher amount than Rater 2, we first re-code all of the responses where Rater 1 uses "Human-Like" but

Theme	Code	Keywords
Linguistic Elements	Speaking Style	Accent, List, Articulation, Specific Word Choice
	Prosody	Tone, Inflections, Cadence, Pitch, Monotone, Raspy, Emotion
	Disfluency	Pauses, Filler Words
	Speed	Fast, Slow, Rushed
External	Quality	Background Noise, Microphone, Recording, Clipping
	Liveliness	Breathing, Mouth Noises, Nasal
Intuition	Human-Like	Natural, Human
	Robotic	Robotic, Glitchy, Mechanical

Table 3: Our codebook for categorizing responses from participants in our user study. We analyze each response using eight unique codes with corresponding keywords, then group those codes into three major themes.

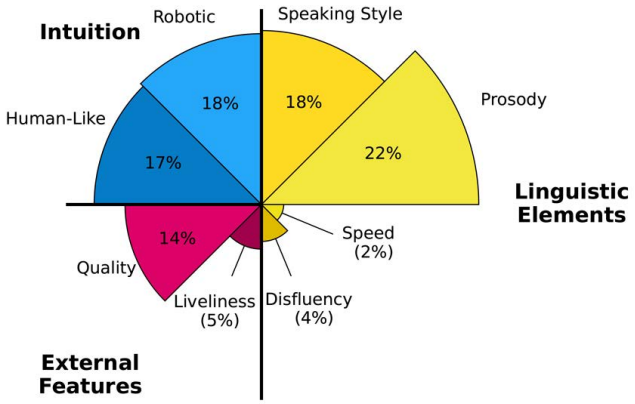

Figure 5: Appearance rate of the eight codes used in the thematic analysis. We designate codes contained within each theme with similarity in color. It demonstrates that Prosody is the most common factor that contributes to classification decisions by people, while Speed is the least common.

Rater 2 does not, then move to the responses where Rater 1 uses "Robotic". During the re-coding process, Rater 3 chooses between the two codes used by the original raters and changes the code that was not chosen to match. The re-coding continues until a strong agreement is met. In total we manually code all $n = 24,240$ responses, remove $n = 3,237$ bad responses, and achieve a resulting Cohen's Kappa coefficient of $\kappa = 0.82$ on the $n = 21,003$ remaining. We label each response with 1-2 codes (1 for rater agreement; 2 for disagreement) totaling $n = 24,987$ codes. We use this number as the denominator in our descriptive statistical tests.

5.2 Reasoning Themes

While manually coding the $n = 24,240$ responses, three themes emerged in the codes for participant reasoning. We separated the themes based on whether the participants referred to a Linguistic Element, talked about an External Feature outside of the speech, or seemed like they were relying on Intuition (e.g., experience, feeling, or guessing). The distribution of codes and appearance rates can be seen in Figure 5.

Linguistic Elements. The largest theme is Linguistic Elements, which is comprised of the Prosody, Speaking Style, Speed, and Disfluency codes. The majority of participants try to find some kind of fault in the voice and choose to believe the sample is human if they do not find any faults.

Prosody ($n = 5,553$; 22%) is the most commonly referenced topic among all of the responses. This is primarily when people identify problems or naturalness in tone and varying degrees of cadence (e.g., *"It has a dull tone."* (P697), *"The speech felt like it was too "perfect" with the timing between words to be natural"* (P439)). Participants also state that emotion is a key factor within Prosody in believing a sample is human-generated (e.g., *"The speech is very enthusiastic, emotion is more of a human trait"* (P1202)).

Disfluency ($n = 924$; 4%) and Speed ($n = 560$; 2%) are the two least present codes among the responses. These generally focus on the existence of pausing and the speech being too fast or too slow. For example, P1056 states that in one of the samples *"the pausing was very jerky and unnatural"* and P1078 believes that one of their samples was *"a human that speaks really fast."*

Speaking Style ($n = 4,518$; 18%) contains some of the most unique and detailed responses of the Linguistic Element codes. These responses contained many reasonings based on things that the participants do not believe computer-generated voices are capable of. For example, many participants do not believe computers can generate accents (e.g., *"I do not believe I've ever heard a computer generated voice with a proper English accent"* (P877), *"The accent in her voice was distinct making her seem human. I would say when someone has a heavy accent or impediment it makes it seem more human"*(P525)). One participant (P626) states, *"The accent gives it away. I have never heard a quality [computer-generated] voice with a believable accent,"* and later reconsiders their stance, commenting *"The Irish accent makes me think it's human, although I'm beginning to think my stance on accents (meaning human) might be incorrect."* Another participant also states that *"I don't think a computer can mimic a speech disorder"* (P610). We note that these preconceptions persist despite the multitude of accent options in popular voice assistant programs [44, 62].

Another trend in Speaking Style is trying to dissect the way the phrase or parts of the phrase are stated. For example, P556 comments that *"People do not say, on November twenty-two"* and P1132 believes that the way the person in the sample was speaking *"sounded like a snarky answer which is human."*

Overall, when people are reasoning using Linguistic Elements, they are pulling from their experiences interacting in various types of conversations. They look for deviations from their experiences or their perceived capabilities in computer voice generation and use that as the basis for determining if a sample was human or computer-generated.

Finding 1. *Participants have pre-conceived ideas of what computer voice generation is capable of which impacts how they reason about detecting deepfake audio.*

External Features. Outside of the voice, some people look for faults in features of the audio sample or look for sounds outside of

the speech to influence their decision. External Features, containing the Liveliness and Quality codes, is the smallest of the themes but gives insight into some artifacts that people key in on when making discriminations on audio samples.

Liveliness ($n = 1,121$; 5%) indicates that something in the audio suggests to the participant the presence of a person. The most common liveliness feature was the appearance of breathing, which overwhelmingly convinces people that the audio is human generated. For example, P254 indicates that breathing means human, saying, *"I could hear breathing and that made it sound human,"* while P1203 looks for lack of breathing saying, *"They sound like they could be robotic or human. If I could hear him breathe it would be obvious it was a man."* Note that both participants were incorrect in their choices based on these reasonings.

Quality ($n = 3,556$; 14%) refers to various issues associated with the sample that alert the participant to problems. The most common term that participants use with this code is "background noise" (e.g., *"The background noise felt like a static computer noise and I think it may be a computer"* (P641), *"there is distortion and some sort of high frequency noise in the background"* (P757)). Some participants reference the recording equipment, such as P605 saying, *"It sounds slightly robotic, but it's hard to tell if it's a computer or just microphone feedback."* Others reference distortion or audio clipping saying things such as, *"there was a lot of stutter and audio clipping at the end that made her sound very robotic"* (P892). Some participants also use background sounds to make their decisions, such as P856 saying, *"This was easy. I clearly hear laughing in the background, so that tells me this is being recorded live and is a human voice"* and P64 saying, *"There was someone else talking in the background."*

The External Features theme demonstrates that just generating quality speech is not enough for deepfakes since people look at all parts of the audio sample when making decisions. The presence of additional artifacts can help humans accurately discriminate, while also giving deepfake generators a way to influence recipients by adding artifacts that they look for when deciding an audio sample is human generated.

Finding 2. *Audio artifacts play a key role in how participants discriminate on deepfake audio which could easily be manipulated by deepfake generators.*

Intuition. While the other two themes are more straightforward, some responses were more difficult to label since the reasonings were not detailed. The Intuition theme, containing the codes for Human-Like and Robotic, represents reasons that the participants either could not articulate or narrow down.

Robotic ($n = 4,437$; 18%) is the code given to responses that insinuate the participants believe the sample was computer-generated. These responses use terms such as robot, machine-like, and unnatural. Generally, the participants do not identify specific traits, saying things like, *"it has a very robot and computer-generated sound"* (P74).

Human-Like ($n = 4,318$; 17%) conversely references the belief that the sample is human without specifics. Responses with this code refer to things such as "natural", "real", or "human-like". For example, P275 claims that the sample *"sounded like how a real person would talk"* while P7 says, *"Just basic instinct. Seemed pretty natural"*.

We see that over a third of the codes come from people making instinctive responses in their decisions. Not everyone knows what they are keying in on, but it is important to know how often people determine the audio source based on general impressions.

Finding 3. *While not as prevalent as linguistic features, participants still heavily rely on intuition when discriminating on deepfake audio.*

Knowing the frequency of certain types of responses and the general themes is just the start in understanding how people discriminate audio samples and begins to answer **RQ2**. A further dive into how these mindsets affect performance and how different types of deepfakes perform will give a better understanding and contextualize the performance of our study participants.

5.3 Thematic Reasoning by Data Type

We now split the codes by their true classification value (i.e., real audio vs deepfakes) and share our results in Figure 6. This figure exhibits three trends to explore, based on three groupings of codes. The first are the codes (e.g., Human-Like, Robotic, Liveness) which exhibit unbalanced results based on data type. For example, the Robotic code represents a unique case in which false positives outweigh true negatives for real samples, suggesting that while the code is a strong factor in detecting fake samples, it can also lead to heavy false positives if relied upon too heavily. The other two imbalanced cases, namely Human-Like and Liveliness, represent the only cases in which false negatives outweigh true positives. Since the goal of a deepfake adversary is to produce a false negative sample, this result suggests that they are incentivized to focus on these two qualities during the generation phase.

The next group, consisting of Disfluency, Speed, and Quality, represents the case in which humans exhibited their strongest average performance for both data types (i.e., correct classifications heavily outweigh incorrect classifications). This grouping provides evidence that humans perform well on samples primarily exhibiting sentence mistakes, odd speed, and quality issues.

The final group (e.g., Prosody, Speaking Style) represents cases where results are more balanced, mainly for fake samples. Humans identifying these codes perform closer to random guessing when the sample is fake, signaling that how the sentence is spoken can be a strong factor for adversaries to focus on, albeit not as reliable as Human-Like or Liveliness.

Finding 4. *Humans missclassify fake samples which exhibit organic features and real samples that sound robotic at high rates. Also, humans perform well on real and fake samples that primarily feature sentence mistakes, odd speed, and quality issues.*

5.4 Confounding Variables

During our study, we observe participant behavior that could have a minor effect on the results. Since the participants received the audio in a survey setting, a small number of participants insinuated that their decision was based on the belief that they were intentionally being tricked. For example, P131 stated that they were *"not sure. Sounded human but I'm expecting to be tricked "* and P308 felt as if

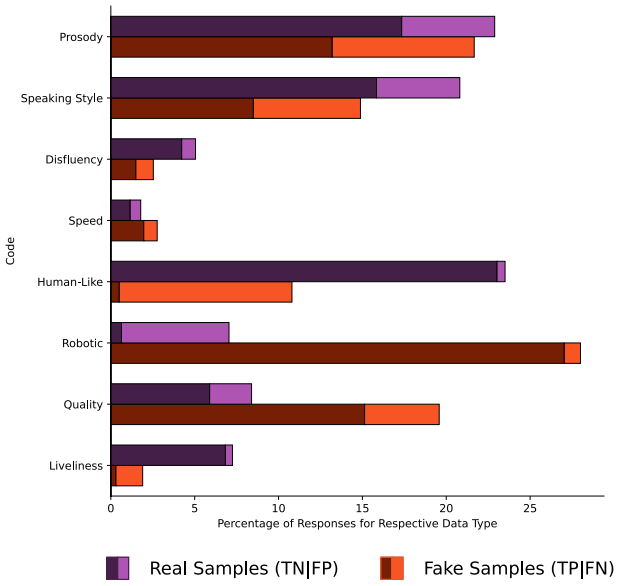

Figure 6: The percentage of code distributions for each sample type (e.g., fake, real), both adding up to 100%. Real samples are represented as true negatives (TN) and false positives (FP) while fake samples are represented as true positives (TP) and false negatives (FN). While Prosody is the most balanced of the codes between the data types (i.e., both real and fake audio have a high number of classifications using Prosody), people perform the best on samples they classify using Disfluency, Speed and Quality.

"the oddities I hear are less because of AI and more because of intentionally manipulated audio". This mindset could be a byproduct of the survey setup that does not translate to real-world applications.

While our survey setup is designed to have each audio sample be discriminated on separately, factors between samples sometimes influenced the participants' decisions. This was particularly noticeable with the Wavefake dataset when participants used the repeated voice to make their decision since it consists of only one speaker (e.g., *"I'm basing my decision solely on the fact that in the previous examples, the voice is the same"* (P452) and *"It sounded similar to the last one, now I'm second guessing myself"* (P1113)).

Additionally, we noticed participants often explain away the artifacts inside the deepfake audio. In some cases, people thought that the deepfakes were human with a filter or bad recording, (e.g., *"sounds like it could have been read by a human, but with a filter placed over top of the recording"* (P1092) and *"there were issues with the sample like someone recorded it"* (P1142)). In both cases, the participants mislabeled deepfakes as human-generated. Some people also explained monotone, unemotional speech in a context that they were familiar with (e.g., *"everything except 'immediately' sounds exactly what I'd expect a news reporter to say"* (P930)).

Some of the decisions that were made by participants were based on content, which is unique to the dataset. For example, P686 said they *"think the fact that the voice was giving biographical information made it seem more human than computer"*. While we did not have control over this variable of the deepfakes, it demonstrates that what is being said can be just as important to the success of a deepfake as the quality.

Finding 5. *Many additional factors impact the way humans classify including a distrusting environment, recently heard audio for comparison, audio content, alternative reasoning for faults and audio sample construction (e.g., length)*

These distributions in codes, overall (Figure 5), by data type (Figure 6) and the confounding variables, help us understand why people are making certain classifications and the emerging themes behind them. The combination of these analyses addresses **RQ2**.

6 COMPARISON WITH ML DETECTORS

We train several models in order to compare our user study results to model performance. In this section, we describe our model training process and compare the results of the average human to the average model to answer our **RQ3**.

6.1 Model Training Process

We train four widely used baseline audio deepfake detectors: three baseline models (RawNet2, LFCC-LCNN, CQCC-GMM [15, 77]) and the SSL-wav2vec2.0 XLS-R-based detection model [60] which is, to our knowledge, the best-performing model on ASVspoof2021. Each model is retrained exclusively using one of the three datasets defined in Section 3.1 while the other two datasets are untouched in training and testing. We repeat this process for all three datasets and models creating a total of 12 models. For the ASVspoof2021 data, we train with the training set exclusively, to not overlap with the evaluation data we selected for the user study. For WaveFake and FakeAVCeleb, the datasets do not provide a train/test split, therefore we remove the audio files (x) that we use in the survey and retrain the models with the remaining data ($D_{total} - x$). Each model trains for the default 100 epochs and follows the training pipeline detailed in their associated GitHub repositories. Since previous work [2] shows that there exists randomness in training models, we provide all 12 models in our Zenodo for reproducibility of our results.[3]

6.2 Detection Models vs. Human Performance

While the extrapolated accuracy calculated in Section 4.3 explains how well the users classify on each dataset, we now focus on the human's performance on the sampled audio, D, to perform a deeper comparison against the models' performance using a sample to sample direct comparison. Specifically, we identify whether models and humans are misclassifying the same audio samples or different ones, how this varies between real and deepfake audio, and whether the trend is consistent across all of the datasets. Across the four models, we calculate the average performance on D, the average model, to compare against the humans.

Of the 3,400 samples in D_{asv}, humans attain 76% accuracy compared to the average models' accuracy of 78%. Figure 7 shows the classification breakdown as sets between the human model and the average model (e.g., a set would be the group of audio samples that both the human model and average model correctly named as

[3]https://zenodo.org/doi/10.5281/zenodo.11044486

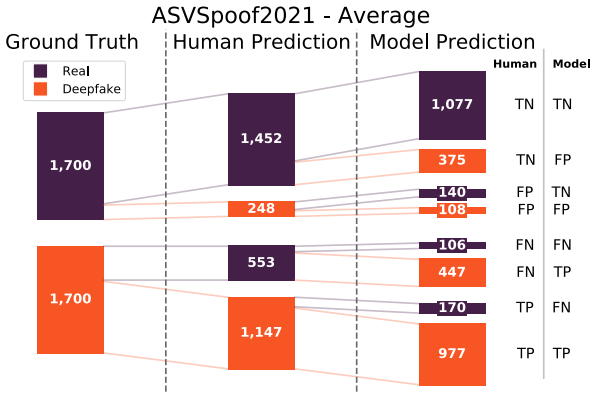

Figure 7: The classification breakdown for the average human and average model performance on the ASVspoof2021 samples D_{asv}. We show that the average human is more prone to false negatives while the average model is more prone to false positives.

	Real		Fake	
	χ^2	p-value	χ^2	p-value
Wavefake	33.67	<0.001	121.63	<0.001
ASVspoof2021	96.24	<0.001	122.40	<0.001
FakeAVCeleb	208.58	<0.001	0.05	0.25

Table 4: χ^2 test comparing the accuracy of the average person to the ML model accuracy for real and deepfake audio samples. We see that there is a significant difference in the way that humans and models classify audio samples for each dataset except for the FakeAVCeleb fake audio.

real known as the TN/TN set). Interestingly, humans and models correctly agree approximately 60% of the time with 28% agreed true positive and 32% agreed true negative, and both humans and models missed approximately 6% of D_{asv} evenly balanced between false positives and false negatives. We see that humans are more prone to false negatives while models are more prone to false positives. For D_{wave}, humans classify with a 70% accuracy and models classify at 74%. The humans and models correctly agree 51% of the time with an even split between true negatives and true positives. Both the models and humans misclassify 8% of the dataset. In contrast to D_{asv} and D_{wave}, humans outperform the models in D_{celeb} with 86% accuracy compared to the models accuracy of 74%. The models and humans agree 64% of the time, and both misclassify 3% of D_{celeb}. Classification breakdown for D_{wave} and D_{celeb} can be seen on our companion website.

Additionally, we use a χ^2 test to see whether there is a significant difference in the way that humans and models classify real and deepfake audio. Table 4 shows the results of the χ^2 tests for each of the real and fake portions of D for each of the datasets. Except for the fake portion of D_{celeb}, we observe that there is a significant difference in the way that people and models classify samples for each data type. This contradicts the notion that models simply classify better than people, but rather there is a difference in the way that each classify audio sample, which needs to be considered when characterizing defense strategies against deepfakes.

Regardless of the detection mechanism, there are on average 6% of the total sampled audio per dataset that both the humans and models misclassify. This is evenly split between real and deepfake audio samples. Humans are more susceptible to false negatives and deepfake models are more susceptible to false positives, showing that improvements are required in mechanisms to be applicable in real-world scenarios.

Finding 6. *Models do not strictly perform better than humans, but rather there is a significant difference in the way that humans and models classify audio samples. Humans are prone to false negatives while models are prone to false positives.*

6.3 Thematic Analysis in Models

With measuring the difference between humans and models, we can contextualize the performance of the models using the thematic analysis. We focus on three specific cases to understand what models are missing that humans are not and what both humans and models are collectively missing. We show the three cases that we analyze in Figure 8. We provide the full list of code distributions for each scenario across all datasets on our companion website.

Case (1) We find that of the samples that humans accurately classify as deepfakes and models classify as humans (Human TP, Model FN) the two biggest themes are Robotic and Quality. The largest theme is Robotic, and we see that in both cases ASVSpoof had larger proportions of the two themes. Thus, often humans are more likely to identify deepfakes that sound robotic or of poor quality, and models are often missing deepfake samples.

Case (2) Liveliness had the biggest differences between datasets for agreed false negatives. Liveliness accounts for only 0.9% of the themes in FakeAVCeleb where as Wavefake and ASVSpoof are 4.0% and 8.5%, respectively. Comparing within ASVSpoof, we look at Liveliness between the agreed false negative (Human FN, Model FN) and when only humans misclassify the deepfakes (Human FN, Model TP). We see that Liveliness accounts for only 4% of a Human FN and Model TP. Thus, models are more likely to struggle with deepfakes that exhibit Liveliness.

Case (3) The biggest theme that both humans and models struggled with (Human FN, Model FN) is Prosody. Prosody appears 52% of the time in FakeAVCeleb under this scenario. This is the largest single occurrence of a code regardless of classification or agreement, and is nearly double the occurrence of Prosody codes in Wavefake (28%) and ASVspoof2021 (19%).

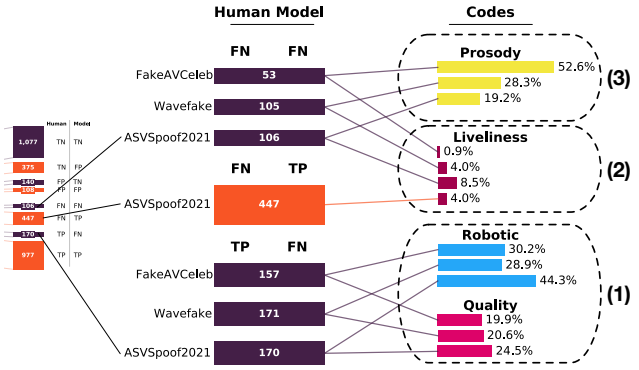

Figure 8: The major contributing codes for three cases: (1) when humans correctly predict deepfakes that models miss, (2) when models correctly predict deepfakes that humans miss, and (3) when humans and models both misclassify deepfakes.

Finding 7. *Humans rely on intuition and recording quality when correctly identifying deepfakes that models miss in D_{asv}. Humans and models both misclassify fake samples at higher rates when the sample is reported to contain Prosody in D_{celeb} and Liveliness in D_{asv}. When contextualized within different scenarios, we observe significantly different distribution behavior between datasets.*

7 DISCUSSION

7.1 Deploying Defensive Pipelines

While model performance is important, optimally deploying a defensive pipeline that considers not only model performance but also human contributions is a difficult problem. As we demonstrated in Section 6, models tend to lean towards a decision of deepfake and thus cause a large number of false positives. If the model is the first line of defense, the recipient may be overwhelmed with warnings leading to threat-alert fatigue [24]. Conversely, if systems rely on people to flag samples and then lean on models for forensic testing, our study showed that humans lead to a large amount of false negatives meaning a larger chance to miss deepfakes. Because both people and models classify audio in different ways and models do not simply directly improve upon human classification, the pipeline of both mechanisms (i.e., humans and models) needs to be considered.

For example, consider an adversary that is targeting a call center. Note that a call center could seek to limit risk to financial crimes for a bank or limit its exposure to automated calls that deplete valuable resources. In the simplest scenario where an incoming call is immediately directed to a call center employee, the model would have to run either concurrently or after the call. Thus, the earliest chance to detect a deepfake happens with the call center employee either independently or with their decisions informed in real-time by a model. In both cases, it is the person who is making the final judgement call on audio. Another complex scenario would include an automated directory that directs the call. A model could be deployed at this stage to determine whether an incoming call is generated audio or not and terminate the call if the model determines the audio to be fake. This scenario allows the model to make the final decision on a potential deepfake. These deployment considerations and pipeline decisions are integral to the audio deepfake detection system, with priorities varying based on the specific scenario (e.g., financial crimes or resource depletion), influencing the system's optimal design.

Additionally, when using models, performance is often described using singular metrics (e.g., accuracy, EER, precision). This limited interpretation of an ML model does not fully capture the nuances of performance and hinders the understanding of how good the models actually perform. ML models should be tested and presented using multiple metrics, including comparisons with human performance, to fully understand what the model is capable of and how best to pipeline it in conjunction with people.

The way forward is a combination of more careful training of models to operate in the space where humans are weak, but also better training of those users. At the current time, both components are simply too inaccurate to form an effective pipeline. Understanding the scenario that the threat of deepfake poses to a system is important and understanding the benefits and limitations of human interactions with deepfakes will allow people to more appropriately deploy the models into real-world settings. Thus, we encourage future work to analyze how systems perform in the context of human interactions with them.

7.2 Human Training

We design our study to minimize the training of survey participants by omitting any feedback on their performance, as shown in Section 4. This approach allows us to achieve a more accurate representation of how humans would perform outside of a survey environment. Therefore, based on our findings, we believe that our qualitative results can offer insight into how humans should be trained in real-world scenarios to more accurately detect deepfakes.

We look at specific factors that contribute to false positive and false negative responses. We learn from **Finding 1** that humans often report that they do not believe a computer can generate accents outside of those found typically in the United States. However, we show that this assumption is false based on cases where participants misclassified audio due to accents and from the existence of additional deepfake audio datasets with a variety of accented audio files [57]. Furthermore, **Finding 2** suggests that the presence or lack of breathing or external background noise can strongly influence human responses toward real or fake, respectively. We observe in the cases of P254 and P1203 that these lead to incorrect classifications. In fact, human-like features cause the largest percentage of false negatives overall. As such, we suggest that training participants be made aware of the capabilities of state-of-the-art deepfake generators and that no single factor within the audio recording be regarded as dominant in influencing classification.

We note here that past studies [47] show that training does not necessarily improve performance but can degrade performance in key areas. In our banking call center scenario, for example, a bank could desire to lower employees' natural skew toward human classification. To achieve this, they could train these employees to use audio artifacts as a dominant factor in reasoning about the

voice they communicate with. Training in this way might reduce false negatives, but at the cost of increasing the false positives, skewing human classification toward deepfakes. This shift limits the availability of the call center by rejecting calls from real customers and still damages the bank in the end. Ultimately, training people to detect deepfakes is a new and complex problem. Our results suggest good focal points for future research to test the best ways to inform people about deepfake capabilities and train them to be better deepfake detectors.

7.3 Reproducibility

Recent work has focused on identifying pathways to reproducible research to encourage better research practices and inspire confidence in results [54]. To promote open and transparent research, we make the coded data, trained models, and code to run our figures publicly available in our linked Zenodo. Additionally, we provide a companion website[4] that centralizes our figures and contains sample audio used in our study. We are unable to provide the direct responses from each person due to the limitations from our IRB. We encourage future work to broaden their threat model and contextualize the performance of their model within the human factors of our data.

8 CONCLUSION

Audio deepfakes are a growing concern not just within the security community, but the broader community worldwide. With recent advancements in ML detection of deepfakes, it is not clear whether current mechanisms augment, hinder, or simply contradict human classification of deepfakes. In this study, we analyze how well humans classify deepfake samples, why they make their classification decisions, and how their performance compares to that of ML detectors. To evaluate all of these quantitative and qualitative metrics, we conduct an online user study in which we ask participants to classify samples from the three most cited community audio deepfake datasets as "human" or "computer-generated". Our findings suggest that humans achieve an average accuracy of 73% on samples from these datasets, with notably improved performance on real samples. Furthermore, we identify what factors lead them to their decision (e.g., prosody, accents, background noise) and compare how the impact of these factors changes between data types (i.e., real or fake). We also compare performance between humans and ML models, demonstrating that models do not strictly perform better than people, but rather there is a significant difference in the way that humans and models classify audio. These results provide us the beginning of how best to approach training humans to become better audio deepfake detectors and better contextualize the performance of humans and ML models. To promote reproducibility in this field, we provide all of our results and our survey structure online.

ACKNOWLEDGMENTS

We would like to express our gratitude to our sponsors for enabling this research. This work was supported by the US National Science Foundation grant CNS-1933208 and grant CNS-2206950, and the Office of Naval Research grant ONR-OTA N00014-21-1-2658.

[4]https://sites.google.com/view/better-be-computer/home

REFERENCES

[1] 2023. Resemble-Ai/Resemblyzer. Resemble AI.
[2] Hadi Abdullah, Kevin Warren, Vincent Bindschaedler, Nicolas Papernot, and Patrick Traynor. 2021. SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems. In *Proceedings of the IEEE Symposium on Security and Privacy*.
[3] Alan Agresti and Christie Franklin. 2009. *Statistics: The Art and Science of Learning from Data*. Upper Saddle River, NJ.
[4] Ehab A AlBadawy, Siwei Lyu, and Hany Farid. 2019. Detecting AI-Synthesized Speech Using Bispectral Analysis. In *CVPR Workshops*.
[5] Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Liwang Lu, and Zhenguang Liu. 2023. Transferring Audio Deepfake Detection Capability across Languages. In *Proceedings of the ACM Web Conference*.
[6] Daniel J Benjamin and James O Berger. 2019. Three Recommendations for Improving the Use of p-Values. *The American Statistician* 73, sup1 (2019), 186–191.
[7] Logan Blue, Luis Vargas, and Patrick Traynor. 2018. Hello, Is It Me You're Looking For? Differentiating Between Human and Electronic Speakers for Voice Interface Security. In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*.
[8] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. 2022. Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction. In *Proceedings of the USENIX Security Symposium*.
[9] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. 2022. Voice-Face Homogeneity Tells Deepfake. *arXiv* (2022).
[10] Akash Chintha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. 2020. Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection. *IEEE Journal of Selected Topics in Signal Processing* 14, 5 (2020), 1024–1037.
[11] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv* (2017).
[12] Beilin Chu, Weike You, Zhen Yang, Linna Zhou, and Renying Wang. 2022. Protecting World Leader Using Facial Speaking Pattern Against Deepfakes. *IEEE Signal Processing Letters* 29 (2022), 2078–2082.
[13] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Proceedings of the ISCA Interspeech Conference*.
[14] Davide Cozzolino, Matthias Nießner, and Luisa Verdoliva. 2022. Audio-Visual Person-of-Interest DeepFake Detection. *arXiv* (2022).
[15] Héctor Delgado, Nicholas Evans, Jeeweon Jung, Tomi Kinnunen, Ivan Kukanov, Kong Lee, Xuechen Liu, Hye-jin Shim, Md Sahidullah, and Hemlata Tak. 2021. ASVspoof 2021 Baseline CM and Evaluation Package. https://github.com/asvspoof-challenge/2021.
[16] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 2020. RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.
[17] Marwa Elpeltagy, Aya Ismail, Mervat S. Zaki, and Kamal Eldahshan. 2023. A Novel Smart Deepfake Video Detection System. *International Journal of Advanced Computer Science and Applications (IJACSA)* 14, 1 (2023).
[18] Chao Feng, Ziyang Chen, and Andrew Owens. 2023. Self-Supervised Video Forensics by Audio-Visual Anomaly Detection. *arXiv* (2023).
[19] Christian Canton Ferrer, Brian Dolhansky, Ben Pflaum, Joanna Bitton, Jacqueline Pan, and Jikuo Lu. 2020. Deepfake Detection Challenge Results: An Open Initiative to Advance AI. https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/.
[20] Emily Flitter and Stacy Cowley. 2023. Voice Deepfakes Are Coming for Your Bank Balance. The New York Times - https://www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html.
[21] Joel Frank and Lea Schönherr. 2021. WaveFake: A Data Set to Facilitate Audio DeepFake Detection. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.
[22] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake Detection by Human Crowds, Machines, and Machine-informed Crowds. *Proceedings of the National Academy of Sciences* (2022).
[23] Ammarah Hashmi, Sahibzada Adil Shahzad, Wasim Ahmad, Chia Wen Lin, Yu Tsao, and Hsin-Min Wang. 2022. Multimodal Forgery Detection Using Ensemble Learning. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.
[24] Wajih Ul Hassan, Shengjian Guo, Ding Li, Zhengzhang Chen, Kangkook Jee, Zhichun Li, and Adam Bates. 2019. NoDoze: Combatting Threat Alert Fatigue with Automated Provenance Triage. *Network and Distributed Systems Security Symposium* (2019).
[25] Joe Hernandez. 2023. That panicky call from a relative? It could be a thief using a voice clone, FTC warns. National Public Radio - https://www.npr.org/2023/03/22/1165448073/voice-clones-ai-scams-ftc.

[26] Hafsa Ilyas, Aun Irtaza, Ali Javed, and Khalid Mahmood Malik. 2022. Deepfakes Examiner: An End-to-End Deep Learning Model for Deepfakes Videos Detection. In *Proceedings of the International Conference on Open Source Systems and Technologies (ICOSST)*.

[27] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. 2023. AVFakeNet: A Unified End-to-End Dense Swin Transformer Deep Learning Model for Audio–Visual Deepfakes Detection. *Applied Soft Computing* 136 (2023), 110124.

[28] Keith Ito and Linda Johnson. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset.

[29] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. In *Advances in Neural Information Processing Systems*.

[30] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2019. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. *arXiv* (2019).

[31] Piotr Kawa, Marcin Plata, and Piotr Syga. 2022. Attack Agnostic Dataset: Towards Generalization and Stabilization of Audio DeepFake Detection. In *Proceedings of the ISCA Interspeech Conference*.

[32] Piotr Kawa, Marcin Plata, and Piotr Syga. 2022. Defense Against Adversarial Attacks on Audio DeepFake Detection. *arXiv* (2022).

[33] Piotr Kawa, Marcin Plata, and Piotr Syga. 2022. SpecRNet: Towards Faster and More Accessible Audio DeepFake Detection. *arXiv* (2022).

[34] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S. Woo. 2021. Evaluation of an Audio-Video Multimodal Deepfake Dataset Using Unimodal and Multimodal Detectors. In *Proceedings of the Workshop on Synthetic Multimedia - Audiovisual Deepfake Generation and Detection*.

[35] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. 2022. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*.

[36] Pavel Korshunov and Sébastien Marcel. 2020. Deepfake Detection: Humans vs. Machines. *arXiv* (2020).

[37] Seth Layton, Thiago De Andrade, Daniel Olszewski, Kevin Warren, Carrie Gates, Kevin Butler, and Patrick Traynor. 2024. Every Breath You Don't Take: Deepfake Speech Detection Using Breath. *arXiv preprint arXiv:2404.15143* (2024).

[38] Seth Layton, Tyler Tucker, Daniel Olszewski, Kevin Warren, Kevin Butler, and Patrick Traynor. 2024. SoK: The Good, The Bad, and The Unbalanced: Measuring Structural Limitations of Deepfake Datasets. In *Proceedings of the USENIX Security Symposium (Security)*.

[39] Sangjun Lee, Donggeun Ko, Jinyong Park, Saebyeol Shin, Donghee Hong, and Simon S. Woo. 2022. Deepfake Detection for Fake Images with Facemasks. In *Proceedings of the Workshop on Security Implications of Deepfakes and Cheapfakes*.

[40] Meng Li and Xiao-Ping Zhang. 2023. Robust Audio Anti-Spoofing System Based on Low-Frequency Sub-Band Information. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

[41] Haoxin Ma, Jiangyan Yi, Chenglong Wang, Xinrui Yan, Jianhua Tao, Tao Wang, Shiming Wang, and Ruibo Fu. 2022. CFAD: A Chinese Dataset for Fake Audio Detection. *arXiv* (2022).

[42] Kimberly T. Mai, Sergi D. Bray, Toby Davies, and Lewis D. Griffin. 2023. Warning: Humans Cannot Reliably Detect Speech Deepfakes. *PLOS ONE* 18, 8 (2023), e0285333.

[43] Hafiz Malik. 2019. Securing Voice-Driven Interfaces Against Fake (Cloned) Audio Attacks. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*.

[44] Taylor Martin. 2018. How to Change Your Google Home's Voice. https://www.cnet.com/home/smart-home/how-to-change-the-voice-of-google-home/

[45] Mary L. McHugh. 2012. Interrater Reliability: The Kappa Statistic. *Biochemia Medica* 22, 3 (2012), 276–282.

[46] Microsoft. 2023. Speaker Recognition | Microsoft Azure. https://azure.microsoft.com/en-us/products/cognitive-services/speaker-recognition.

[47] Jaron Mink, Licheng Luo, Natã M. Barbosa, Olivia Figueira, Yang Wang, and Gang Wang. 2022. DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks. In *Proceedings of the USENIX Security Symposium*.

[48] Daniel Mas Montserrat, Hanxiang Hao, S. K. Yarlagadda, Sriram Baireddy, Ruiting Shao, János Horváth, Emily Bartusiak, Justin Yang, David Güera, Fengqing Zhu, and Edward J. Delp. 2020. Deepfakes Detection with Automatic Face Weighting. *arXiv* (2020).

[49] Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena. 2015. All Your Voices Are Belong to Us: Stealing Voices to Fool Humans and Machines. In *Proceedings of the European Symposium on Research in Computer Security (ESORICS)*, Günther Pernul, Peter Y A Ryan, and Edgar Weippl (Eds.).

[50] Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does Audio Deepfake Detection Generalize?. In *Proceedings of the ISCA Interspeech Conference*.

[51] Nicolas M. Müller, Karla Pizzi, and Jennifer Williams. 2022. Human Perception of Audio Deepfakes. In *Proceedings of the International Workshop on Deepfake Detection for Audio Multimedia*.

[52] Ajaya Neupane, Nitesh Saxena, Leanne Hirshfield, and Sarah Bratt. 2019. The Crux of Voice (In)Security: A Brain Study of Speaker Legitimacy Detection. In *Proceedings of the Network and Distributed System Security (NDSS) Symposium*.

[53] Amy Neustein. 2010. *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*. Springer Science & Business Media.

[54] Daniel Olszewski, Allison Lu, Carson Stillman, Kevin Warren, Cole Kitroser, Alejandro Pascual, Divyajyoti Ukirde, Kevin Butler, and Patrick Traynor. 2023. "Get In Researchers; We're Measuring Reproducibility": A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[55] Alessandro Pianese, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. 2022. Deepfake Audio Detection by Speaker Verification. In *IEEE International Workshop on Information Forensics and Security (WIFS)*.

[56] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. In *Proceedings of the International Conference on Machine Learning*.

[57] Ricardo Reimao and Vassilios Tzerpos. 2019. FoR: A Dataset for Synthetic Speech Detection. In *Proceedings of the International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*.

[58] Emma Roth. 2022. James Earl Jones lets AI take over the voice of Darth Vader. The Verge - https://www.theverge.com/2022/9/24/23370097/darth-vader-james-earl-jones-obi-wan-kenobi-star-wars-ai-disney-lucasfilm.

[59] Adam Satariano and Paul Mozur. 2023. The People Onscreen Are Fake. The Disinformation Is Real. The New York Times - https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html.

[60] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. Wav2vec: Unsupervised Pre-training for Speech Recognition. *arXiv* (2019).

[61] Sahibzada Adil Shahzad, Ammarah Hashmi, Sarwar Khan, Yan-Tsung Peng, Yu Tsao, and Hsin-Min Wang. 2022. Lip Sync Matters: A Novel Multimodal Forgery Detector. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

[62] Umar Shakir. 2022. Alexa's English Accents Are Now Selectable Without Changing Your Region. https://www.theverge.com/23475121/amazon-alexa-english-accent-switch-update-region-free

[63] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2017. JSUT Corpus: Free Large-Scale Japanese Speech Corpus for End-to-End Speech Synthesis. *arXiv* (2017).

[64] Chengzhe Sun, Shan Jia, Shuwei Hou, Ehab AlBadawy, and Siwei Lyu. 2023. Exposing AI-Synthesized Human Voices Using Neural Vocoder Artifacts. *arXiv* (2023).

[65] Jenny Tang, Eleanor Birrell, and Ada Lerner. 2022. Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*.

[66] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best Practices for the Human Evaluation of Automatically Generated Text. In *Proceedings of the International Conference on Natural Language Generation*.

[67] Pranshu Verma and Will Oremus. 2023. AI voice clones mimic politicians and celebrities, reshaping reality. The Washington Post - https://www.washingtonpost.com/technology/2023/10/13/ai-voice-cloning-deepfakes/.

[68] Chen Wang, S Abhishek Anand, Jian Liu, Payton Walker, Yingying Chen, and Nitesh Saxena. 2019. Defeating Hidden Audio Channel Attacks on Voice Assistants via Audio-Induced Surface Vibrations. In *Proceedings of the 35th Annual Computer Security Applications Conference*.

[69] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. 2019. VoicePop: A Pop Noise Based Anti-spoofing System for Voice Authentication on Smartphones. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*.

[70] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. 2020. DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices. In *Proceedings of the ACM International Conference on Multimedia*. 1207–1216.

[71] Xin Wang and Junichi Yamagishi. 2023. Investigating Active-Learning-Based Training Data Selection for Speech Spoofing Countermeasure. In *2022 IEEE Spoken Language Technology Workshop (SLT)*.

[72] Xin Wang and Junichi Yamagishi. 2023. Spoofed Training Data for Speech Spoofing Countermeasure Can Be Efficiently Created Using Neural Vocoders. *arXiv* (2023).

[73] Yao Wang, Wandong Cai, Tao Gu, Wei Shao, Yannan Li, and Yong Yu. 2019. Secure Your Voice: An Oral Airflow-Based Continuous Liveness Detection for Voice Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2019).

[74] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y. Zhao. 2021. "Hello, It's Me": Deep Learning-based Speech Synthesis Attacks in the Real World. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[75] Jee weon Jung, Hee-Soo Heo, Hemlata Tak, Hye jin Shim, Joon Son Chung, Bong-Jin Lee, Ha jin Yu, and Nicholas W. D. Evans. 2021. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[76] Ying Xu, Kiran Raja, Luisa Verdoliva, and Marius Pedersen. 2023. Learning Pairwise Interaction for Generalizable DeepFake Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

[77] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. 2021. ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection. In *The Automatic Speaker Verification and Spoofing Countermeasures Challenge*.

[78] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. 2023. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE Transactions on Information Forensics and Security* 18 (2023), 2015–2029.

[79] Hong Yu, Zheng-Hua Tan, Zhanyu Ma, Rainer Martin, and Jun Guo. 2018. Spoofing Detection in Automatic Speaker Verification Systems Using DNN Classifiers and Dynamic Acoustic Features. *IEEE Transactions on Neural Networks and Learning Systems* 29, 10 (2018), 4633–4644.

[80] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing Your Voice Is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*.

[81] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. VoiceLive: A Phoneme Localization Based Liveness Detection for Voice Authentication on Smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*.