SoK: Deepfake Audio Detection

*Note: Sub-titles are not captured in Xplore and should not be used

Amelia Matheson
Computer Science Undergraduate
University of Arizona
Tucson, Arizona
ameliamatheson@arizona.edu

Jose Chavez
Computer Science Undergraduate
University of Arizona
Tucson, Arizona
josechavez@arizona.edu

***Abstract*—Many AI-generated tools are being utilized to mimic human voices known as Audio Deepfakes. Additionally, other digital tools allow audio files to be edited. Audio manipulation detection is becoming a main concern for the justice and forensics field. The ability to detect audio manipulation is necessary to maintain reliable evidence. There are two main kinds of audio manipulation, multiple- source and single-source manipulation. Multiple manipulations use various forms to forge audio. An example is using audio splicing to move small portions of audio and AI-generated audio to append new audio to change the message completely. Single-source manipulations are defined as using one source to alter the audio. In short, this paper will discuss the current accuracy of audio manipulation detection and the methods to detect it. It also informs scientists of possible vulnerabilities in detecting methods and how to improve on weaknesses. Currently, audio manipulation detection takes advantage of machine learning with logistic regression to create a classification for audio. This research paper will examine major discoveries and structures for further analysis to improve detection accuracy by analyzing weaknesses and offering effective solutions.**

***Index Terms*—Audio Manipulation Detection, Logistic Regression, Fake Voice, Machine Learning**

I. Introduction

Problem statement: The problem we want to solve is how can we create an effective detection process to detect audio forgery and manipulation. Another issue we will explore is how can we improve current detection processes. Our research questions are how can the accuracy and efficiency of audio manipulation detection be improved, specifically for imitation and synthetic deepfakes, and what are the current limitations of detecting manipulated audio using machine learning models in real-world situations?

Significance: The significance of solving this problem is immense as audios have been and continues to be exploited in a vast range of fields. Audio deepfakes have been used in the forensic field and in the court room. For example, a UK child custody case was putting a father at danger of losing his child's custody. The mother used audio and video deepfakes to make the father appear aggressive and dangerous to allow a child to be near him [7]. Fortunately, digital forensics experts were able to examine the audio and videos. They concluded that the audio and video evidence had been manipulated by deepfake software. Furthermore, false audio can lead to incorrect judgments, which could have serious consequences for individuals, families, and society. Furthermore, individuals could face wrongful consequences. In Maryland, a fake recording of a principal saying antisemitic and racist comments was spreading through the Baltimore county. Shortly, it was discovered by government officials that it was in fact an audio deepfake. The principal was forced on leave and police guarded his house from potential violence [2]. Humanity will benefit by being able to produce correct decisions from audio that has not been tampered with. Since the sophistication of Audio Deepfakes have advanced, we have seen an increase of misinformation being spreaded. Attackers have generated audio deepfakes of government officials and politicians to modify public opinions for defamation and propaganda. The biggest case of audio deepfakes occurred when a CEO was misled into transferring over $243K to an attacker, they used real time audio deepfake to make this happen [14]. Not only are criminals targeting political figures but wealthy individuals. Audio deepfakes have caused confusion and nearly wrongful consequences to innocent individuals. As discussed before, a father nearly lost his relationship with his child, a principal almost had their career and reputation ruined, and a CEO transferring over $243K to criminals. To mitigate the negative impact of audio deepfake misuse, we aim to examine existing methods and identify under-researched areas, with the goal of improving both model performance and detection techniques. Detecting AI-generated audio will prevent wrongful decisions and maintain safety.

Existing literature: Current work has identified at least three major types of audio deepfakes, replay-based, imitation-based, and synthetic-based deepfakes. Replay-based deepfakes are when a target's voice is recorded to be replayed at later time with malicious intent. Imitation-based deepfakes refer to when an audio is modified to imitate another speech. Synthetic-based deepfakes are audios generated based of raw audio samples then read aloud a given speech [4]. Current work to detect audio forgery has been using machine learning with logistic regression to classify audios. The accuracy is at 0.98 as of right now [11]. Proposed by Kumar-Singh and Singh [1], their proposed machine learning model is a Quadratic Support Vector Machine model (Q-SVM). Compared to other models such as Linear SVM and weighted K-Nearest Neighbors (KNN), the Q-SVM exceeded 97.56% compared to the other models. We will discuss detection models more concerned with imitation-based and synthetic-based deepfakes. Two major attributes to take into account are generalization and stabilization of detecting methods. Improving generalization will allow models to more accurate detections and stabilization is more concerned with the consistent detections. Currently, there are ways to testing a model's generalization and stabilization by using an "Attack Agnostic Dataset" [5], which is a dataset carefully crafted from three major datasets to determine a model's generalization and stabilization. However, there are more issues for audio deep fakes. One of the issues is able to detect audio deep fakes with background noise or voices with accents [1]. Not to mention, large datasets are difficult to access. There are very few methods that are able to accurately detect deepfakes that are training with small sized datasets. Currently, one method was produced to detect audio deepfakes in an efficient time range with a relatively accurate rate considering it trains with a small dataset [8]. While there is not much material about self-supervised learning and audio deepfake detection, there does exist some research. Currently, there is a self-supervised spoofing audio detection model (SSAD) [12]; however, it has yet to have similar results as it performs weaker than other models. On the other hand, other developing models that are taking advantage of both self-supervised learning and machine learning. Semi-supervised models are useful when there is limited labeled data and an abundance of unlabeled data. Unfortunately, the limited quantity of labeled data can cause the semi-supervised model to be skewed and produce biased results. Hybrid models take advantage of different models or methods to produce more accurate results. Earlier this year, a new hybrid model combines a SSL model with an advanced classifier to result with a promising detection model [3]. The SSL is WavLM, it is used as a frontend feature extractor. After raw audio samples are fed into WavLM, the model creates a layers of frames of embedded features. Then frame layers of embedded features are fed into the classifier. The hybrid model uses a Multi-Fusion Attentive classifier which combines the output from different layers and steps of the output of the WavLM. This allows features to be distinctly extracted and finds the most informative features. In order to achieve a high spoofing rate, the hybrid model requires predetermined parameters and pre-trained weights. Furthermore, it can be computational expensive when to detect audio deepfakes. .There is an investigation of creating hybrid models of self-supervised models with deep learning models. Further, research is needed to create a more robust model to detect audio deep fakes. The current progress in self-supervised has been recently explored with improving accuracy results, but requires increased hardware requirements [9].

Solution sketch: This problem's solution is not quite out of grasp, but it is a significant undertaking to find. Current detection methods rely on large-scaled machine language and deep learning models that are difficult to manage and build. A possible solution lies in "Self-Supervised Learning", which makes use of unlabeled input data to train models. This method is hypothesized to be more efficient as well as scalable. To implement such a solution, we need to collect a large amount of unlabeled training data and teach our model to accurately detect fake audio without supervision. Another possible solution is using hybrid models. By combining different methods, it could lead to higher accuracy rates. This would include examining different methods and identifying their weaknesses and strengths. This would allow us to find efficient hybrid models by merging different aspects of the models.

Challenges: Current literature points out that the "Self-Supervised Learning" method is not yet been widely tested for detecting fake audio. One study that did implement it reported low accuracy. So our main challenges in this project are to effectively investigate the SSL method, decide how best to train ML and DL models under this framework, and efficiently integrate SSL into the fake audio detection arena. To obtain a sufficient quality SSL model, we would need large datasets to properly train the model to perform adequately. Furthermore, when trained properly, the SSL model would be manageable and perform than a Semi-Supervised model and other models. This being said, current works have struggled obtaining large enough datasets to train models. Additionally, current works have indicated that ML models are capable of high accuracy with the cost of time. When deciding which features to use to create an effective, it becomes time-consuming as it is done manually and very labor intensive. There has been research comparing the generalization of hand-crafted selections of features against deep learning choosing its own features. The generalization of the hand-crafted feature selection performed poorly compared to the deep learning features [13]. It is reported that the classification procedure is time-consuming despite only two seconds and 270 features. The ML is not easily scalable as audios can be longer, it will take even more time to classify. DL models excel as features are chosen by the model itself. Unfortunately, there has not yet been a DL model with high accuracy compared to the ML model. As stated before, current detection methods have issues when detecting audios with significant background noise or accents. Additionally, criminals have applied to post-processor to avoid detections due to some methods depending on patterns of latent noise. We will need to take to investigate how we can decrease the effectiveness of these factors.

II. Definition of Terms

In this section, basic terminology related to topics presented within this paper are explained.

A. Machine Learning Models

Machine learning (ML) model is a program which finds patterns and decides based on a previous dataset. ML models are trained with training data to improve accuracy, then are tested with testing data. The testing data has never been seen by the model, the ML model will get tested by how accurate it can correctly classify the data. The model's accuracy will consistently improve through learned data [6].

B. Deep Learning Models

Deep Learning (DL) is a subset of of machine learning which itself is a subset of artificial intelligence and statistics. While machine learning depends on pre-determined features, deep learning decides which features to use. Often, deep learning is related to neural networks with various layers of neurons which allow the deep learning model to complete more complex problems [10].

C. Logistic Regression

Logistic regression predicts the probability of an object belonging to one of two classes. Using a logistic function to transform predicted values between zero and one. It has been applied in market research, medical analysis, and banking. Logistic regression finds the relationship between the input and the target variable [6].

D. Support Vector Machines

Support Vector Machines are another type of machine learning that is used to classify objects. It finds the hyperplane which best divides objects from two different classes. Commonly, separate binary classification training is needed. [6].

III. Threat Model

Although we will not be designing the ML and DL models necessary for fake audio detection ourselves, it is important to consider the environment these models are expected to operate in. Ultimately, our security goal is to accurately and efficiently detect audio deep fakes, especially those that contain background noise and accented voices. Therefore we must assume that bad actors that produce these fakes are using tactics like theses to make the deep fakes sound as authentic as possible by closely imitating real audio. Defenders can include financial crimes analysts (in the event of extortion or ransom), missing persons detectives, media personalities whose identities and ideas can be misrepresented with audio deep fakes, etc. In any situation where audio deep fakes can be used, someone will need to defend against them.

IV. Methodology

Our proposed solution includes analyzing current detection procedures and examining what components could be improved upon to create more effective detections. One example would looking into what machine learning features are needed to determine accurate Audio Deepfake detections. As mentioned before, looking into how accents and background noise affects our detection. After finding how these factors affect our detection, we will explore how to minimize the severity of the factors mentioned. We will understand how detection models work and if there are any dependencies. If there are any dependencies, we can investigate if they can be exploited and avoided. As mentioned before, bad actors found that some detection models depend on latent noise patterns. They would apply post-processor, such as a low-pass filter, to avoid detection. Understanding how a model detects audios will allow us to find any possible vulnerability.

V. Evaluation

Criteria for success: To measure our project success, our criteria will be based off how generalizable our solution is, efficiency of the model with our solution implemented, improvements against current audio deepfake issues such as background noise,any relative improvements of accuracy, and how feasible our solution is to implement.

Datasets: We plan to use existing datasets to help us understand the scope of the problem, as well as what elements of audio deep fakes our solution needs to target. Specifically, we will be using the FoR dataset to understand our dataset and comprehend the robustness of it as well. The FoR dataset is used to test the accuracy of model's detection. Not to mention, we will use HAV-DF, a Hindi audio deepfake dataset. This will help to develop a more comprehensive model.

Timeline

The first step is to investigate current known methods for detecting audio deep fakes; this means that we will also need to determine the weaknesses of these methods so that we know what needs to be improved upon. Weaknesses would include not only accuracy rate, but as well as efficiency, generalization, and the severity of possible vulnerabilities. After finding any possible weaknesses, we need to investigate how an actor could benefit from these weaknesses and how they could avoid them. As stated before, poorly structured models would depend on latent noise patterns. We estimate that this will take about a week (following the submission of this proposal). Next, since we have settled on further exploring SSL, we estimate that significant time should be allocated to synthesizing information about how to train models on unlabeled data, identifying its limitations, and evaluating its cost-effectiveness (time, data, resources, etc). Not to mention, explore any current work with SSL to find any common issues. Additionally, identify if there are any additional components that improve SSL. A possible option is creating a hybrid model by adding another model to the SSL. However, a hybrid model would be an external goal. Our main focus is the SSL, this may take us three to four weeks. And lastly, we need to connect SSL to audio detection by applying this method of ML/DL to the audio detection process. As there is not much material available that discusses this integration, we will need to come up with a plan for integration ourselves. We will attempt to gather much research on this integration to properly integrate the system. We will use the remainder of the time (4-5 weeks) to do this as well as to write our final report.

VI. Review of Academic Sources

In this section, we will take a look at existing literature and discussion on Audio Deepfakes (AD). Many studies have been conducted evaluating the efficacy of various detection methods as well as the robustness of audio datasets used to train and test models.

A. A Review of Modern Audio Deepfake Detection Methods by Zaynab Almutairi and Hebah Elgibreen

In their study, Almutairi and Elgibreen explain that Audio Deepfakes were initially designed to enhance our lives by way of audiobooks and Text-to-Speech tools. However, they have been widely used to spread misinformation, manipulate public opinion through propaganda, defame individuals, exploit money from unsuspecting victims, and even to commit acts of terrorism. Because of this myriad of threats, it is becoming increasingly imperative for cybersecurity specialists to develop effective ways of detecting deepfakes. The general process of detection includes first preprocessing each audio clip to be analyzed, and transforming them into Mel-spectograms (audio features). The features are then inputted into the ML or DL detection model for training and testing, after which the audio clip is classified as fake or real.

There are two types of ADs: imitation-based and synthetic-based deepfakes. Imitation-based deepfakes transform speech so that it sounds like the target audio. The original and target audios are recorded with similar characteristics and then a masking algorithm transforms the original audio's signal to say the speech in the target audio. Imagine using the president's voice in the original audio to recite a public message outlined in the target audio. Synthetic-based deepfakes (aka Text-to-Speech (TTS)) transforms text into natural-sounding human speech in real time using a text analysis model, an acoustic model, and a vocoder. Clean and structured audio is collected, then the TTS model is trained on the collection. A transcript of the audio and the target speaker's voice is given to the generation model; the two models and the vocoder work together to generate the audio deepfake file.

ML detection models are effective for the most part, achieving 98% success rate with a Logistic Regression model [11]. However, the input audio data must be manually preprocessed to extract the features necessary for training the model. Thus, most ML detection methods suffer from a inefficient, time-consuming preprocessing step that oftentimes introduces inconsistencies. DL detection eliminate this step and instead analyzed audio that first is transformed to an image. For example, Deep4SNet visualized the audio and classified it based on a 2D CNN histogram model with 98.5% accuracy. But it isn't scalable. Furthermore, other DL models report varying success rates, some of them being highly inaccurate. Both ML and DL models require the input data to be processed and/or transformed in special ways before the models can be trained.

B. Deepfake Audio Detection via Feature Engineering and Machine Learning by Farkhund Iqbal, et.al.

Iqbal and the four other authors of this paper provide a brief note on one attack that audio deepfakes are routinely used to carry out: replay attacks. A replay attack is simply when the target speaker's voice is replayed to the audience or victim through a device. One way this can be done is through a far-field detection system, in which a "far-field microphone recording of the target has been repeated on a phone handset with a loudspeaker" [4]. Another way is through a cut-and-paste detection system: short recordings are spliced together to form full sentences aimed at miscommunicating information.

Iqbal et.al. also propose their own approach for detecting Audio Deepfakes, using the Fake or Real (FoR) dataset for their experiment. As a first step, they extract audio features from the data, similar to what ML models already do in the preprocessing step. They identified that 270 features can be extracted from the audio samples in the dataset, but narrowed this number down to 65 relevant features useful for training the models. Next, they normalize the features through data framing. They take an audio sampling value after every second in a clip and set the sampling rate. Then, the number of frames is defined as the product of the sampling rate and the duration of the clip. The frequency domain of the audio signal is visualized on a Mel power spectrogram. Their study results revealed that the SVM, MLP, and XGB models were most accurate on the for-2sec audios in the dataset, achieving 97.57%, 94.69%, and 94.52% respectively [4].

C. Towards generalisable and calibrated audio deepfake detection with self-supervised representations by Octavian Pascu, et.al.

Pascu and the four other authors of this paper provide a brief review of how models can be tested for their generalization and calibration. Generalization is the model's ability to sufficiently perform on new data and produce accurate results. Their research proceeds to explain how current models struggle with new data after training, due to the lack of sufficient training data. Using twelve models, all of models performed relatively good. However, the models were tested on the ASVspoof' 19 dataset, which they considered a simple dataset. When the models were tested on a challenging dataset, the In-the-Wild dataset, all of the models features needed to be fine tuned in order to produce quality accuracy rates. By fine tuning the features, the models lack any generalization.

Pascu et.al. points out how current research regularly addresses how classifiers, in detection methods, are given a property to measure their calibration magnitude. While this aspect is well researched, calibration of classifiers are not researched enough.

They also propose their own model to improve generalization and calibration performance. Using a self supervised model with partial pretrained representation , then training a binary classifier on the representations. The representations came from the wav2vec method. Using the wav2vec method allows Pascu et.al. to identify factors. After the initial setup, they used the ASVspoof'19 dataset to train their model. Comparing to Raw2Net, one of the best models concerning generalization, their proposed model performed %8.8 equal error rate (EER) while Raw2Net performed %30.9 EER. There is a notable difference between both models, with their proposed significantly improved. However, Pascu et.al. do mention they it does require greater computational power but it is still reasonably attainable.

VII. Limitations and Vulnerabilities

The landscape of audio deficit detection research has a few limitations and vulnerabilities. First, current studies have done limited testing on audio deep fix with background noise or accented speakers. Although deep learning models are very accurate in detecting a specific subset of audio deepfakes, when faced with audios containing background, real world noise, their performance suffers [1]. This is most likely due to the fact that most of the data sets that these models are trained and tested on do not include many audio files with real world noise. Another possible reason is that the models are simply not extracting these features so they cannot use them to help detect fake audio. In th is case, more robust models are required. Ask for languages, English prevails as the most tested language in detection processes. There has been some research of Arabic deepfakes. However, most datasets contain just one dialect or form of Arabic while there exists many around the world. More diverse data sets are definitely required if we wish to properly defend against fake audios.

Moreover, the common detection models are not scalable at all. Most of the audio clips in widely tested datasets are only about 2 to 10 seconds long. So the models are very inefficient for detecting audio deepfakes that are any longer than this, considering that they consume a lot of resources as is. More complex and layered audio deepfakes need to be included in the testing and training process so that the models we are using in real life situations can accurately detect maliciously created audios.

VIII. Our Findings

Our research of this problem allowed us to answer one of our questions: what methods can be implemented to make fake audio detection more efficient and accurate? We learned that self-supervised learning models offer practical solutions to many of the problems that plague status-quo machine learning and deep learning models used for audio detection. Namely, self-supervised learning models avoid issues of scalability and time-consuming pre-processing that plagues common ML and DL models. Most if not all ML and DL models are trained and tested on very short audio clips. They are largely ineffective and inefficient for detecting fake audios that are any longer than that; but these are exactly the audios that we encounter in the real world. Self-supervised learning models offer practical use cases that help defend us against audio deep fakes we are more likely to come across. These include political propaganda audios, audio that is meant to mislead the victim and so that they can be extorted or exploited, and audios that misrepresent the victims actions or beliefs.

A promising SSL was evaluated by Oneata et. al. They chose a model from the wave2vec 2.0 family models, One that accepts 2 billion parameters as input. This model performs ”unsupervised pre-training on audio clips which allow them to self learn useful speech representations without the need of phonetic and linguistic annotations”. After testing on a multitude of datasets, the pre-trained, SSL model achieved 8.8% error compared to RawNet2’s 30.9%. Other SSL models from Wav2Vec2 and WavLM also perform well, reducing error to 5.1% and 8.6% respectively with 300M parameters [9].

IX. Further Research

As for further research and development, we propose that professionals in this field who are dedicated to detecting audio deepfakes should further refine detection procedures as well as examine which components can be improved within the machine learning and deep learning algorithms. As we discovered while researching this problem, machine learning and deep learning models must either be provided features to perform on or extract them themselves. Therefore, it can be of benefit to audio deepfake detectors to further analyze which features these models are taking into account. That way, they can further pin down which elements within audio deepfakes are more identifiable.

Secondly, we believe it is important for these professionals to understand which features are recurrent across deepfakes as well as which features of authentic audio are being exploited by bad actors. Criminals and attackers are becoming increasingly sophisticated in their methods; if they learn which features machine learning and deep learning models are targeting, this gives them the ability to better circumvent some of these detection points. So, it is imperative for professionals to think like attackers and learn which elements within authentic audio are easily replicable.

Like in other cyber-security areas, audio deepfake criminals are becoming quite crafty. For example, they have the ability to apply post-processors to the audio deep fakes in order to prevent audio detection. They do this through low-pass filters, which only transmit audio signals of a frequency that are lower than a selected cutoff frequency. So, it is very possible that they can guess which frequency ML and DL models are tested on and apply filters to the fake audio so that it has an easier time sneaking past detection methods. Another thing attackers can possibly exploit are languages that are widely spoken. Take English and Spanish in the United States, for example: we hypothesize it is easier to generate fake audio in these languages because attackers have plenty of authentic audio material on which to train their deepfake generators. And lastly, is it possible for attackers to try adding emotion signifiers or other human-like tendencies in their audio? Little things like chuckles and sniffles would definitely make detection a little bit more difficult and would require models to extract more features besides tone, pauses, and pitch.

X. Conclusion

We hope that by investigating audio deep fakes, known ways of detecting them, and synthesizing a new method for accurately thwarting deep fake attacks, we will be gain a better understanding of how to defend against sophisticated social engineering security threats. As well as understanding and analyzing current model detections, we hope to find to ways to improve detection models. By implementing our solutions, we aspire to see current issues, such as background noise and accents, to be less of an issue to generate more accurate results. In addition, we hope that our proposed solution will provide a net positive to the security world as bad actors become increasingly skilled in exploiting human interactions for illicit gains.

REFERENCES

[1] Almutairi, Zaynab, Elgibreen, and Hebah. A review of modern audio deepfake detection methods: Challenges and future directions. In *Algorithms*, volume 15, 2022.

[2] B. Finley. Deepfake of principal's voice is the latest case of ai being used for harm, Apr 2024.

[3] Y. Guo, H. Huang, X. Chen, H. Zhao, and Y. Wang. Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12702–12706, 2023.

[4] F. Iqbal, A. Abbasi, A. R. Javed, Z. Jalil, and J. N. Al-Karaki. Deepfake audio detection via feature engineering and machine learning. In *CIKM Workshops*, 2022.

[5] P. Kawa, M. Plata, and P. Syga. Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection. In *Interspeech*, 2022.

[6] J. Kufel, K. Bargieł-Łaczek, S. Kocot, M. Koźlik, W. Bartnikowska, M. Janik, Łukasz Czogalik, P. Dudek, M. Magiera, A. Lis, I. Paszkiewicz, Z. Nawrat, M. M. Cebula, and K. Gruszczyńska. What is machine learning, artificial neural networks and deep learning?—examples of practical applications in medicine. *Diagnostics*, 13, 2023.

[7] C. F. Marketing. Deepfake audio evidence used in court to discredit father: Cyfor, Apr 2024.

[8] H. Oiso, Y. Matsunaga, K. Kakizaki, and T. Miyagawa. Prompt tuning for audio deepfake detection: Computationally efficient test-time domain adaptation with limited target dataset. *ArXiv*, abs/2410.09869, 2024.

[9] D. Oneaţă, A. Stan, O. Pascu, E. Oneata, and H. Cucu. Towards generalisable and calibrated audio deepfake detection with self-supervised representations. *Interspeech 2024*, 2023.

[10] M. Paluszek and S. J. Thomas. What is deep learning? In *What is Deep Learning*, 2020.

[11] Rodriguez-Ortega, Yohanna, Ballesteros, and D. Maria. A machine learning model to detect fake voice. In Florez, Hector, Misra, and Sanjay, editors, *Applied Informatics*, pages 3,13, Cham, 2020. Springer International Publishing.

[12] O. A. Shaaban, R. Yildirim, and A. A. Alguttar. Audio deepfake approaches. *IEEE Access*, 11:132652–132682, 2023.

[13] Y. Yang, H. Qin, H. Zhou, C. Wang, T. Guo, K. Han, and Y. Wang. A robust audio deepfake detection system via multi-view feature. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13131–13135, 2024.

[14] L. Yasur, G. Frankovits, F. M. Grabovski, and Y. Mirsky. Deepfake captcha: A method for preventing fake calls. *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, 2023.