The Ethical Dangers and Detection Challenges of Deepfake and Synthetic Audio

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Robert Jason Hudson

Spring 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Richard D. Jacques, Department of Engineering and Society

Introduction

Recent advancements in artificial intelligence have led to significant developments in audio manipulation technologies, particularly in the realm of synthetic audio, sometimes also referred to as “Deepfakes.” This technology is capable of creating highly realistic sound recordings that are often able to accurately imitate the speech patterns of others. Notably, it has made headlines within popular culture, as social media and music streaming services have begun to see an increase in artificially generated audio content (Henkin, 2023). While this seemingly innocuous usage has already been deemed controversial in certain discussions, what I found to be more pressing was the dangers behind it. Now that synthetic audio generation has become widely accessible, opportunities for unethical misuse have become rampant across the globe.

In one prolific example, the technology was used to mimic the voice of a UK-based energy company's CEO, convincing an employee to transfer over $200,000 into a fraudulent account (Harwell, 2019). But stealing money is only one consequence – the misuse of synthetic audio represents a potent threat to personal lives and society at large. Scammers can easily exploit these technologies to impersonate individuals, fabricate statements, and conduct sophisticated frauds, typically resulting in severe consequences. The realistic nature of synthetic audio makes it particularly effective and dangerous, as it can potentially be difficult to distinguish from authentic audio, especially vulnerable groups like the elderly who tend to be more trusting in digital communications. A recent study also found that the average individual is not perfect when it comes to distinguishing between a genuine voice and its deepfake counterpart in multiple trials of major languages (Mai et al, 2023).

Fortunately, efforts have been made in both research and politics to curb dangerous misuse and hold malevolent users accountable. The American government’s Defense Advanced

Research Projects Agency has already created its own research program known as Semantic Forensics (SemaFor) in an effort to create robust algorithms to detect fabricated media and plan to commercialize them as soon as this year (DARPA, 2024). Another campaign, the Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVSpoof), is a bi-annual event organized by researchers from across the world in an effort to further advance research in detection and safety surrounding synthetic audio (Yamagishi et al, 2021).

While many experts in the field are actively addressing the threats posed by synthetic audio through numerous initiatives, the average person may still lack the essential knowledge needed to prevent falling victim to such scams. While systems that allow users to evaluate audio snippets exist, these are often inaccessible to the general public and even unintuitive to some users. It is my goal with this paper to inform others about the vast number of risks posed by the accessibility of synthetic audio creation, explore the considerable challenges involving its detection for the average person, and discuss the strategies currently being implemented and plans for the future to address this rapidly increasing problem. I also plan the possible steps that can be taken to make detection more accessible and increase awareness of this dangerous issue.

I: Synthetic Audio Creation

The technologies behind synthetic audio generation have transformed from niche experimental systems only accessible to scientists and researchers to becoming powerful tools capable of producing highly realistic and convincing sound bites that millions around the globe have access to. Many models that create cloned audio are fully accessible to the public, sparking interest and even temptation. But how do these models work?

To generate realistic synthetic audio, especially voices that mimic human speech, developers begin by gathering substantial amounts of audio recordings. The sound data is then

used to train deep learning models on the nuances of human speech, including tone, pitch, and inflections. The training phase is particularly important since it allows an AI to learn and replicate the unique characteristics of a voice, making the synthetic audio nearly indistinguishable from real human speech (Boreilli et al, 2021). The more audio of a person's voice that is available, the more realistic and lifelike the synthetic creations become.

However, recent advancements by OpenAI, a non-profit largely responsible for the recent shift towards artificial intelligence's mainstream coverage, have seemingly reduced the amount of data required to just fifteen seconds of audio in order to clone a voice (Davies, 2024). The specific technology behind its "Voice Engine" is unknown and its usage is restricted to very few due to the companies fears of misuse, but its existence proves that synthetic audio creation will only get easier as time passes.

Various AI models are able to generate synthetic audio, most notably convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs are known for accurately completing tasks correctly that require recognizing patterns in data (Mehrish et al, 2023). In the context of synthetic audio, they analyze the spectral features of audio signals and process the distinct aspects of sound by examining the frequency and amplitude of audio samples. This makes CNNs ideal for tasks like voice conversion, where the goal is to alter an existing voice into another while maintaining a natural sound.

RNNs, on the other hand, are known for their ability to remember previous inputs in a neural network's internal state. This feature is useful for speech synthesis since RNNs can predict what the next sound segment will be based only upon the sequences they have already processed (Mehrish et al, 2023). This also allows them to generate speech that flows naturally and has more range than a typical synthesized voice.

Beyond these neural network technologies, synthetic audio generation sometimes uses another synthesis technique known as Text-to-Speech (TTS) technology. TTS systems convert written text into spoken word, an increasingly common tool used in various applications ranging from virtual assistants to real-time communication aids. Modern TTS systems leverage AI to improve the natural sound and flow of synthetic voices, allowing them to sound realistic and more human-like (Mehrish et al, 2023).

While there are other audio generation techniques that exist, these are the primary ones that drove my research efforts because they represent the more accessible and well-documented approaches. This information was not only used to better inform myself on their development but also contributed to my technical project wherein I propose my own creation of a synthetic detection algorithm.

II: Synthetic Audio Misuse

The creation of accessible synthetic audio technology has come with a downside a lot of danger and proliferation of illegal activity. The FTC recently published data on how "consumers reported losing more than $10 billion to fraud in 2023, marking the first time that fraud losses have reached that benchmark. This marks a 14% increase over reported losses in 2022" (2024). They go on to state that the increase is due to multiple technological factors, with one of them in particular being voice phishing technologies.

Bad actors using this technology often use psychology to manipulate their victims in tandem with this deepfake audio. An increasingly common occurrence is a tactic known as an emergency scam. This is a situation where a perpetrator will create a scenario that leverages the urgency and emotional distress of their victim in order to acquire financial or personal information. This scam evolves further with synthetic audio, since scammers can use a fake

voice of a family member claiming to be in dire need of money for an emergency, preying on the vulnerable groups and exploiting their concern for loved ones. This approach is not only highly effective, but it also inflicts a considerable amount of emotional distress on the victim.

Similarly, synthetic audio can be used to facilitate identity theft and extortion. Scammers have the ability to mimic an individual's voice to authorize fraudulent transactions, access confidential information, or manipulate individuals into revealing sensitive data like login credentials or account numbers. This not only has the potential to affect the average individual but also large-scale companies that do not have the appropriate safeguards put into place. This type of fraud leads to substantial financial losses for both people and banks, which in turn can erode the trust people typically put into financial institutions.

But the consequences of this technology are not just monetary losses. Even if someone safeguards themselves from scams, their reputation is still at risk of being tarnished. Recently, a high school athletic director was caught using AI to create a fabricated audio clip that impersonated his school's principal wherein racist remarks were said (Singer, 2024). The fake recording went viral online and as a result, the principal faced unwarranted backlash and was put on administrative leave. Even after being proven innocent, it is possible an event such as this has a tangible impact on a person's reputation and personal self-worth, with possible lifelong negative effects. This is just one example of a reputation being ruined; it is possible other incidents such as this have occurred, and the perpetrators have not yet been caught.

Even then, it is possible to consider what else is at risk on a larger scale: possibly the fabric of our society and the governmental institutions we trust in. While no actual incidents have occurred yet, synthetic audio poses a direct threat to the United States' legal system by creating fake evidence to obstruct justice. In one scenario, audio recordings could be manipulated to

create fabricated statements or confessions that are then used in legal proceedings. This hypothetical has also created a scenario known as “the liar’s dividend” wherein someone accused of a crime can claim evidence like a verbal confession or video proof to be artificially generated (Schiff et al, 2023). This would not only cause havoc within individual legal proceedings, but it also has the potential to put strain on entire legal systems due to the possibility of prolonged cases where defendants claim evidence is fabricated.

The world’s political stages are also potentially susceptible to the dangers of synthetic audio. Deepfake audio can be used to create speeches or statements that can mislead the public or tarnish the reputation of political figures. Notably, during the 2020 US election cycle, there were concerns about deepfakes being used to influence voter behavior by spreading misinformation or fake endorsements (Schiff et al, 2023). Another instance this year saw a fake audio clip of a political candidate go viral just before an election, wherein his digital double claimed he rigged an election and planned to intentionally raise grocery prices on his constituents (Lyngaas et al, 2024). The political applications of this technology are a threat to our nation’s democracy and the existence of free and true speech.

All of the examples provided are recent, within the past five years. But the number of incidents reported is only going to increase with time, which makes efforts to detect and prevent these situations all the more urgent.

III: Detection Challenges: Technological and Accessibility

As synthetic audio technologies advance, distinguishing between genuine and manipulated audio becomes increasingly difficult. While current algorithms are fairly successful, with some of the best models coming out of the most recent ASVspoof challenge ranging from 92% to 99% accuracy, is it possible the rapid development of new artificially intelligent

technologies renders these systems obsolete (Yamagishi et al, 2021). That potential obsolescence means research needs to adapt alongside advancements in an effort to keep up and hopefully even outpace technological developments.

It is obvious that human senses are not the solution to this problem, as we can sometimes not trust our senses. In fact, a recent study determining if humans could identify synthetic versus real audio recordings had concerning findings where humans failed somewhat often. They finish the report, stating: “it raises the question whether humans would be less able to detect deepfake speech created using the most sophisticated technology available now and, in the future,” (Mai, et al, 2023). Humans are only going to get worse at detecting synthetic audio so now is the time to start familiarizing ourselves with ways to detect it besides just using our ears.

The detection technology is not foolproof either, however. A large component that contributes to the accuracy of machine learning models is their dataset. If the training data lacks diversity or fails to include samples manipulated by more recent synthetic audio techniques, the system's ability to recognize newer types of manipulated audio is likely to diminish. This necessitates ongoing updates to the datasets and algorithms powering these detection systems. This can be done by integrating a wide range of voice samples and manipulation methods to stay effective, but due to the sheer amount of different languages, dialects, and accents, no model can ever be fully perfect.

Another barrier is the excessive cost of developing and maintaining sophisticated detection systems. Simply put, a localized system is unaffordable for many individuals and smaller organizations. Moreover, the complexity of these systems and specialized knowledge required to operate them poses a significant challenge when considering a wide adoption. This complexity not only makes it difficult for non-experts to use these tools but also limits their

deployment in environments where they are most needed. Even consumer-friendly online systems offering simple and easy detection methods are stuck behind paywalls and subscription models that most organizations and individuals are not going to see the worth in paying for. The general public already lacks an awareness of synthetic audio risks, making them unlikely to even consider using such a service in the first place. Without widespread understanding of the potential threats posed by synthetic audio, the demand for these tools remains low which in turn slightly hinders their development. Fortunately, development has not been fully halted and strides are being made every day in protecting those most at risk and furthering cracking down on synthetic audio misuse.

IV: Technical and Legal Efforts to Curb Misuse

The recent uptick in synthetic audio related incidents has prompted significant responses from various groups, including high-level research initiatives like the U.S. Government DARPA's Semantic Forensics (SemaFor) program. This program represents a concerted effort to develop tools and algorithms that are capable of distinguishing real content from manipulated or synthesized media. The agency's investments in the field of synthetic media detection have spawned the creation of numerous analytical methods and technologies aimed at detecting, attributing, and characterizing synthetic media. It was recently announced that SemaFor is entering its final phase, meaning DARPA is encouraging broader engagement from commercial industries and academic researchers to use their advancements and continue the momentum in combating manipulated media.

To facilitate this collaboration, DARPA has launched two significant initiatives. The first is an analytic catalog of open-source resources developed under SemaFor, which provides researchers and industry professionals with access to innovative tools for media verification. The

second initiative, the AI Forensics Open Research Challenge Evaluation (AI FORCE), focuses on advancing machine learning models that can effectively differentiate between authentic images and those generated or altered by AI. This initiative will feature a series of challenges that encourage participants to innovate and refine their existing detection models (DARPA, 2024).

Besides technical efforts for deploying accessible systems, governments worldwide are beginning to recognize the potential dangers posed by synthetic audio technologies and have begun enshrining protections against it into law. For instance, in the United States, legislators have proposed bills specifically targeting deepfake technologies. One proposal, the Protecting Consumers from Deceptive AI Act, seeks to criminalize the malicious creation and distribution of deepfake content. This act requires that creators of synthetic media disclose their manipulation by embedding digital watermarks and metadata within the content, providing consumers with the information necessary to discern real from fake media (Rep. Eshoo, 2024).

Similarly, the European Commission has included synthetic audio in its broader regulations on artificial intelligence. The AI Act proposed by the Commission classifies deepfake technologies as high-risk applications, requiring strict compliance with transparency standards (E.U., 2024). These include clear disclosures when synthetic media is presented, ensuring that individuals are aware when the content they are consuming is not entirely authentic.

These proposals are a start, but legislation has a long way to go if we wish to co-exist with these artificial technologies. We must put human safety and well-being above technological advancement, and while we wait for our government to take initiative to protect us, it is imperative awareness of the issue is spread in the meantime.

V: Increasing Awareness and Accessibility

Increasing awareness through education is fundamental in combating the misuse of synthetic audio. Public awareness campaigns can be facilitated through partnerships with media outlets, social media platforms, and public institutions to disseminate information about the dangers of synthetic audio scams. These campaigns should highlight real-world cases of fraud and manipulation, providing clear examples of the threats and offering guidance on how to respond if individuals suspect they are being targeted by a scam.

It is also necessary we call upon our government representatives to enact more comprehensive regulations on the technology, restricting its usage for malicious reasons while still protecting our freedom of expression. Regulatory bodies should enforce transparency from companies developing synthetic audio and visual technologies, mandating them to disclose the capabilities of their technologies and what measures, if any, they have in place to prevent their misuse. This would include the implementation of detailed ethical guidelines and review boards to oversee the development and deployment of new artificially intelligent generative technologies.

It is also crucial we develop and distribute user-friendly detection interfaces (web and mobile applications) that enable users to easily evaluate and identify synthetic audio accurately. This could be achieved through government funded initiatives for both academic institutions and private companies, allowing them to collaborate on innovative solutions for public benefit. It is essential that this detection software remains publicly accessible and free as well. This would ensure those who need it most and are at high-risk of these attacks are able to use it freely and easily.

Conclusion

While the development of synthetic audio technologies offers groundbreaking advancements in media, they also pose potential risks that could be exploited for fraudulent purposes. Throughout this paper, we have discussed the complexities of synthetic audio creation, the many ways it can be misused, and the substantial challenges in detecting such manipulations. Moreover, we have explored some of the technical and legislative efforts aimed at curbing this misuse and proposed comprehensive strategies to enhance awareness and improve the accessibility of detection technologies.

My research has shown that synthetic audio can be a powerful tool when used ethically but requires vigilant oversight to prevent its abuse. Recent regulatory measures and public awareness campaigns are critical tools to help individuals and institutions acquire the knowledge needed to recognize and combat synthetic audio scams.

The development of accessible detection technologies remains paramount. Investment in research and the sharing of these technologies can empower the public to defend against audio-based fraud effectively. Initiatives like DARPA's Semantic Forensics program reveal a potential for collaborative efforts between government, industry, and academia to advance this cause.

Ultimately, the path forward must involve a balanced approach that fosters innovation while safeguarding against the ethical and security risks posed by synthetic audio. By increasing societal awareness, strengthening regulatory frameworks, and enhancing technological defenses, we ensure that the benefits of synthetic audio are realized without compromising the integrity and security of digital communications. This comprehensive approach will not only minimize the risks of current synthetic audio technologies but also prepare society for future advancements in this rapidly evolving field.

References

As Nationwide Fraud Losses Top $10 Billion in 2023, FTC Steps Up Efforts to Protect the Public. (2024, February 8). Federal Trade Commission. https://www.ftc.gov/news-events/news/press-releases/2024/02/nationwide-fraud-losses-top-10-billion-2023-ftc-steps-efforts-protect-public

Borrelli, C., Bestagini, P., Antonacci, F., Sarti, A., & Tubaro, S. (2021). Synthetic speech detection through short-term and long-term prediction traces. EURASIP Journal on Information Security, 2021(1), 2. https://doi.org/10.1186/s13635-021-00116-3

DARPA (2024). Deepfake Defense Tech Ready for Commercialization, Transition. Retrieved April 4, 2024, from https://www.darpa.mil/news-events/2024-03-14

Davies, P. (2024) OpenAI unveils AI voice cloning tech but limits availability. (2024, April 1). Euronews. https://www.euronews.com/next/2024/04/01/openai-unveils-ai-voice-cloning-tech-that-only-needs-a-15-second-sample-to-work

Harwell, D. (2019, September 5). An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft. Washington Post. https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/

Henkin, D. (2023). Orchestrating The Future—AI In The Music Industry. Forbes. Retrieved April 11, 2024, from https://www.forbes.com/sites/davidhenkin/2023/12/05/orchestrating-the-future-ai-in-the-music-industry/

High-level summary of the AI Act | EU Artificial Intelligence Act. (2024). Retrieved April 15, 2024, from https://artificialintelligenceact.eu/high-level-summary/

Lyngaas, C. D., Donie O'Sullivan, Sean. (2024, February 1). A fake recording of a candidate saying he had rigged the election went viral. Experts say it is only the beginning | CNN Politics. CNN. https://www.cnn.com/2024/02/01/politics/election-deepfake-threats-invs/index.html

Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. PLoS One, 18(8), e0285333. doi: 10.1371/journal.pone.0285333

Mehrish, A., Majumder, N., Bhardwaj, R., Mihalcea, R., & Poria, S. (2023). A Review of Deep Learning Techniques for Speech Processing (arXiv:2305.00359). arXiv. https://doi.org/10.48550/arXiv.2305.00359

Rep. Eshoo, A. G. [D-C.-16. (2024, March 21). Text - H.R.7766 - 118th Congress (2023-2024): Protecting Consumers from Deceptive AI Act (2024-03-21) [Legislation]. https://www.congress.gov/bill/118th-congress/house-bill/7766/text

Schiff, K. J., Schiff, D. S., & Bueno, N. (2023). The Liar's Dividend: The Impact of Deepfakes and Fake News on Trust in Political Discourse. SocArXiv, Article x43ph. https://ideas.repec.org//p/osf/socarx/x43ph.html

Singer, N. (2024). School Employee Arrested After Racist Deepfake Recording of Principal Spreads. The New York Times. https://www.nytimes.com/2024/04/25/business/deepfake-recording-principal-arrest.html

Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., & Delgado, H. (2021). ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 47–54. https://doi.org/10.21437/