A Sociotechnical Analysis of Deepfake Audio Technology as Evaluated by Machine and Human Observers

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

Vishnu Lakshmanan

Spring 2025

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

Pedro A. P. Francisco, Department of Engineering and Society

STS Research Paper

Introduction:

Artificial Intelligence has made significant advancements in recent years, particularly in the field of voice synthesis. Using machine learning algorithms, AI can now generate highly realistic voice clones that mimic the tone, pitch, and cadence of specific individuals. This technology, while beneficial for applications such as virtual assistants and entertainment, presents a critical threat to security systems that rely on voice authentication. Fraudsters increasingly exploit AI-generated voice clones to bypass identity verification measures, posing a growing challenge for businesses, financial institutions, and individuals.

Modern security systems often incorporate biometric authentication methods to verify users' identities. Among these methods, voice recognition is more popular due to its non-invasive nature and ease of use. Banks, government agencies, and customer service platforms utilize voice authentication to verify identity remotely. This process typically involves analyzing unique vocal characteristics, such as vocal tract shape, pitch, and speaking patterns. However, the rise of AI-driven voice cloning has exposed vulnerabilities in these systems, as fraudsters can replicate these characteristics with high accuracy. Fraudsters use several techniques to exploit AI-generated voice clones for unauthorized access. One common method involves collecting voice samples from publicly available sources, such as social media videos, podcasts, and phone calls. Advanced AI algorithms then analyze these samples to generate synthetic voice replicas. These clones can be used to impersonate individuals during voice authentication processes, enabling unauthorized access to sensitive information, financial accounts, and secure systems. Another technique involves real-time voice manipulation, where fraudsters use AI tools to alter their speech to match a target's voice during live interactions. This

method is particularly dangerous in phone-based authentication scenarios, where the security system must differentiate between genuine and synthetic voices in real time. Additionally, fraudsters may combine voice cloning with social engineering tactics, such as posing as trusted individuals to manipulate victims into disclosing confidential information.

To combat voice cloning fraud, security systems employ various techniques to distinguish between genuine and synthetic voices. One approach involves analyzing the acoustic properties of voice samples, including frequency patterns and background noise. AI-generated voices often exhibit subtle artifacts and distortions that are detectable through advanced signal processing and machine learning models. Another method focuses on detecting anomalies in speech patterns such as unnatural pauses, pitch fluctuations, and irregular intonations. By comparing voice samples to a user's historical data, systems can identify indicators of potential fraud. Additionally, some systems incorporate challenge-response protocols, where users must repeat random phrases or answer spontaneous questions, making it more difficult for fraudsters to replicate specific responses.

Despite these advancements, detecting AI-generated voices remains challenging, as cloning technology continues to improve. Therefore, ongoing research is essential to develop more sophisticated detection techniques, ensuring that voice authentication systems remain reliable and secure in the face of evolving threats. This will allow society to better understand the implications of these technologies and how to utilize them as well as safeguard against the negative implications.

Background/Motivation:

The financial losses associated with AI voice cloning fraud have escalated significantly. Businesses worldwide lost an estimated $35 billion in 2022 due to voice impersonation scams, with cases ranging from unauthorized access to financial accounts to fraudulent business transactions. A high-profile incident in 2020 involved a deepfake voice scam in which criminals impersonated a company executive, leading to a $243,000 wire transfer. As AI voice cloning technology becomes more sophisticated, the financial risks are expected to grow, underscoring the urgent need for enhanced security measures.

Beyond security, AI voice cloning research has diverse industrial applications. In healthcare, synthetic voices are used to assist patients with speech impairments, providing personalized communication solutions. In the entertainment industry, voice cloning enables realistic character dubbing and voice overs, enhancing immersive experiences. Additionally, customer service platforms utilize AI-generated voices to deliver personalized and efficient support. However, the dual-use nature of this technology necessitates careful regulation to prevent misuse while maximizing its benefits.

The rapid advancement of AI voice cloning raises significant ethical concerns. As technology becomes more accessible, the potential for misuse increases, necessitating robust regulations to prevent malicious applications. Ethical considerations include protecting individuals' privacy, ensuring informed consent for voice data usage, and preventing the exploitation of synthetic voices for deception. Moreover, businesses and security professionals must prioritize transparency and accountability when deploying voice authentication systems. Ongoing research is crucial to addressing these challenges and ensuring that security systems can keep pace with evolving threats. By developing advanced detection techniques and implementing

comprehensive cybersecurity protocols, researchers can safeguard sensitive information and maintain public trust. As AI technology continues to scale, collaborative efforts between industry, academia, and regulatory bodies will be essential to strike a balance between innovation and security, ensuring that voice cloning technology benefits society while minimizing its risks.

Methodology:

This study investigates how individuals perceive voice authenticity by examining the relationship between acoustic features, emotional cues, and speech patterns that shape what is considered an authentic voice. Authentic human speech typically exhibits variations in pitch, rhythm, and timbre, along with subtle imperfections such as breathiness, hesitations, and emotional expressiveness. AI-generated voices, despite their growing sophistication, continue to struggle with replicating these nuanced vocal traits, particularly when it comes to emotional spontaneity and contextual variation.

To assess human perception of voice authenticity, a survey was designed in which participants listen to six randomized audio samples and rate each on a sliding scale from “definitely fake” to “definitely real.” The samples include both authentic human voices and synthetic clones, which vary in duration (15s, 30s, 60s) and include both clean and noise-overlaid versions to simulate realistic conditions. AI-generated voices were created using four cloning tools: ElevenLabs, Lovo, FineVoice, and F5-TTS with each tool producing three distinct clones per individual across the three durations. This yielded a total of 336 voice samples in the voice library.

This methodological approach is framed through the Social Construction of Technology (SCOT) framework, with a particular emphasis on the interpretive role of financial institutions as a relevant social group. Financial institutions have emerged as critical stakeholders in the debate over deepfake audio technologies, interpreting these tools not just as technical innovations, but as significant threats to identity verification systems, customer trust, and operational security. This group's perception of deepfake audio as a financial and reputational liability has shaped the urgency behind research on voice authenticity and influenced the design of this study.

The specific inclusion of varied speech durations and background noise was intended to simulate real-world cases in financial fraud scenarios such as impersonated phone calls to banking representatives or automated authentication systems. These features reflect concerns raised by institutions facing rising incidents of audio-based fraud, including voice phishing and unauthorized access to customer accounts. By using randomized logic in the Qualtrics survey (approved by local Institutional Review Board) to vary participant exposure, the study mimics the unpredictability and ambiguity financial employees and systems may encounter when evaluating real versus synthetic voices in practice. This also enables a sufficient number of responses per each question, allowing for large scale statistical data analysis.

This study adopts a hybrid approach by combining evaluations from human respondents with objective "realness" scores from automated tools. This mirrors the reality faced by financial institutions and the need to balance human intuition with automated fraud detection systems. It also aligns with the broader principles of Responsible Research and Innovation (RRI), ensuring that the implications of this research extend beyond technical performance to include institutional priorities, user experience, and systemic resilience.

Literature Review:

Deepfake audio technology, which enables the creation of highly realistic synthetic voices, has introduced significant challenges to cybersecurity. This technology can be exploited to bypass security measures, leading to fraud, identity theft, and misinformation. Understanding this issue through the lens of the Social Construction of Technology (SCOT) framework offers valuable insights into how societal factors influence technological development and its implications.

Deepfake audio technology utilizes advanced machine learning models, such as Text-to-Speech and Voice Conversion, to produce convincing synthetic speech (Li et al., 2024). While these models have legitimate applications ranging from research to industry, their misuse poses serious cybersecurity threats overall. For instance, adversaries can employ deepfake audio to impersonate individuals, thereby deceiving voice recognition systems and humans. This form of attack known as voice phishing, has been on the rise. A notable incident in 2019 involved a con artist using synthetic speech to impersonate a CEO which resulted in a fraudulent wire transfer of around $250,000 (Firc et al., 2023).

Addressing the threats posed by deepfake audio requires robust detection mechanisms. Researchers have explored various approaches to identify synthetic speech. One method involves analyzing the vocal cords. A study demonstrated that deepfakes often model impossible or highly unlikely vocal tract arrangements, and by reconstructing these configurations, it is possible to detect synthetic audio with high precision (Blue et al., 2022). However, deepfake detectors themselves are susceptible to adversarial attacks (Rabhi et al., 2024). Research has shown that state-of-the-art audio deepfake classifiers can be fooled by samples generated using techniques

like Generative Adversarial Networks (GANs). In one study, adversarial attacks reduced a detector's accuracy from 98.5% to nearly 0%, highlighting the need for more resilient detection systems (*Untitled*, n.d.). To enhance the security of detection mechanisms, researchers have proposed content privacy-preserving frameworks. For instance, the SafeEar framework breaks down semantic and acoustic information in audio samples, utilizing only acoustic features for deepfake detection. This approach not only preserves the privacy of the speech content but also maintains high detection accuracy, achieving an equal error rate as low as 2.02% (Li et al., 2024).

Deepfake audio technology exhibits flexibility, meaning different social groups can place various meanings and purposes to it. While technologists and businesses may view it as a tool for innovation by enhancing user experiences through personalized interactions, security professionals and policymakers perceive it as a potential threat to privacy and security. This difference in interpretation influences the direction of technological development and the prioritization of features, such as the balance between usability and security. The development and impact of deepfake audio technology are also influenced by broader societal contexts, including cultural attitudes towards technology, the value placed on privacy, and the level of digital literacy. The developments and innovation caused by this tool might also impact overall legislation in areas where use of deepfake audios technologies is heavily being used. The SCOT framework assumes that technological development is shaped by social, cultural, and political factors, rather than being a purely technical endeavor. Applying this perspective to deepfake audio technology reveals how societal constructs influence both the evolution of the technology and the responses to its associated risks. For instance, societies with high digital literacy may be more resilient to deepfake scams due to better awareness and skepticism. Conversely, in regions

where digital literacy is lower, individuals may be more susceptible to deception, necessitating targeted educational initiatives. Deepfake audio technology presents multifaceted challenges to cybersecurity, enabling sophisticated attacks that can compromise systems and deceive individuals. Through the SCOT framework, it becomes evident that societal factors significantly influence both the development of this technology, and the strategies employed to mitigate its risks. Recognizing the roles of various social groups and the process of interpretive flexibility is crucial in crafting comprehensive responses that encompass technological solutions, regulatory measures, and public education. As deepfake technology continues to evolve, an interdisciplinary approach that considers technical, social, and ethical dimensions will be essential in addressing the complex challenges it presents.

Discussion/Results:

The research question guiding this study is: How does the misuse of deepfake audio technology to bypass security measures relate to the Social Construction of Technology framework? The findings reveal that deepfake audio technology, while originally developed for legitimate applications, has been socially reconstructed into a cybersecurity threat due to its ability to manipulate trust-based authentication systems. The analysis highlights how different social groups interpret and influence the development, regulation, and mitigation of deepfake technology, demonstrating the relevance of the SCOT framework in understanding these dynamics. Findings indicate that deepfake audio presents a growing challenge to cybersecurity due to its use in voice phishing, fraud, and misinformation campaigns. Literature review confirms that while machine learning-based detection methods exist, adversarial attacks significantly undermine their effectiveness. The SCOT framework explains these findings by

illustrating how deepfake audio technology is shaped by competing interests among developers, cybersecurity professionals, policymakers, and the general public. The process of how different social groups view these technologies and eventual closure within these groups affects how society as a broader group responds and adapts to this technology.

Deepfake audio technology enables malicious actors to deceive individuals and security systems, leading to fraudulent activities. For instance, a 2019 case involved a cybercriminal using synthetic speech to impersonate a CEO, resulting in a $250,000 fraudulent wire transfer. This event exemplifies how deepfake audio technology has transitioned from a tool for innovation to a weapon for fraud. Events such as these underscore how certain industries are more sensitive to technologies such as these and must quickly react in order to prevent significant financial losses.

From a SCOT perspective, this shift illustrates interpretive flexibility: developers initially designed deepfake models for accessibility, entertainment, and voice assistance applications. However, hackers and fraudsters redefined the technology for illicit purposes, altering its societal perception from a beneficial innovation to a cybersecurity threat.

Efforts to detect deepfake audio face substantial obstacles. Studies demonstrate that while some detection algorithms achieve high accuracy rates, adversarial attacks can drastically reduce their effectiveness. Research has shown that generative adversarial networks can manipulate voice patterns in a way that evades detection, reducing classifier accuracy to near 0%.

This ongoing challenge highlights the process of stabilization and closure in SCOT. As cybersecurity experts are working to counteract deepfake threats, they influence the technology's trajectory by advocating for more resilient detection methods. In response, adversaries

continuously adapt their techniques, keeping the technology in a state of flux rather than closure. The struggle between attackers and defenders exemplifies the socio-technical co-evolution of deepfake technology and its countermeasures. Regulatory efforts to combat deepfake misuse are still in their early stages. Some governments have proposed criminalizing the malicious use of synthetic media, while some organizations advocate for transparency measures such as watermarking AI-generated audio. The introduction of privacy-preserving detection frameworks like SafeEar, which separates information to maintain detection accuracy while protecting user privacy, represents a step toward stabilization.

The SCOT framework suggests that different social groups influence regulatory responses. Policymakers, responding to public concern and industry lobbying, shape the legal landscape of deepfake audio. Meanwhile, public discourse and media coverage influence how society perceives the urgency of deepfake threats, affecting funding and research priorities. The balance between innovation and security continues to shape how deepfake technology is integrated into legal and technological frameworks. This will be very prevalent in the coming years as the technology will be thrown into the spotlight. The SCOT framework provides a robust lens through which to analyze the social dynamics of deepfake audio. The interpretive flexibility of deepfake technology has led to conflicting uses: one as a tool for accessibility and personalization and also as a cybersecurity risk. Relevant social groups, including developers, cybersecurity professionals, policymakers, and the general public, shape the development and perception of deepfake technology. The ongoing battle between deepfake creators and detection mechanisms illustrates how technological closure remains difficult to achieve, reinforcing the need for solutions that incorporate different stakeholders.

The results demonstrate that deepfake audio technology is not merely a technical issue but a socio-technical construct influenced by various social groups. The SCOT framework highlights the dynamic interplay between technological advancements, societal interpretations, and regulatory measures. As deepfake detection evolves, achieving technological closure will depend on continued collaboration among stakeholders to mitigate security threats while preserving the positive applications of this upcoming technology.

Conclusion:

Deepfake audio technology represents a significant challenge to cybersecurity, with implications extending beyond fraud to misinformation and privacy breaches. This research underscores that deepfake audio is not merely a technical phenomenon but a socially constructed one, influenced by competing interests and interpretations. The SCOT framework highlights how different social groups shape the development, deployment, and countermeasures associated with deepfake technology, revealing a continuous process of evolution rather than resolution.

The key takeaway from this study is that addressing deepfake audio threats requires an interdisciplinary approach. Technical advancements in detection must be complemented by regulatory frameworks and public awareness initiatives. No single solution can fully mitigate the risks posed by deepfake audio, but a combination of robust detection systems and legislative measures is necessary. Future research should explore more resilient detection methods that can withstand adversarial attacks. Additionally, policymakers must work toward clearer regulations that balance technological innovation with security concerns. This should be done hand in hand with those that are developing the technology so that all parties are aware of the risks and consequences while still preserving the ability and incentive to innovate. Another important next step is public education: raising awareness about the existence and risks of deepfake audio

allows individuals to critically evaluate suspicious voice communications. The broader implications of deepfake audio should be examined within the context of digital trust. As AI-generated content becomes more prevalent, society must establish new norms for verifying authenticity. Collaboration between AI researchers, cybersecurity professionals, and legal experts will be crucial in shaping a technological landscape that prioritizes security without stifling innovation. In conclusion, the rise of deepfake audio highlights the need for a proactive and holistic response. By understanding the socio-technical factors driving its development and impact, strategies can be created that not only counteract its misuse but also foster ethical and responsible applications of synthetic technology. The ongoing dialogue between social groups will ultimately determine how deepfake audio is integrated into our digital ecosystems, making it imperative that stakeholders remain engaged in shaping its trajectory.

References:

1. Li, X., Li, K., Zheng, Y., Yan, C., Ji, X., & Xu, W. (2024). *Safeear: Content privacy-preserving audio deepfake detection*(No. arXiv:2409.09272). arXiv. https://doi.org/10.48550/arXiv.2409.09272
2. Firc, A., Malinka, K., & Hanáček, P. (2023). Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon*, *9*(4), e15090. https://doi.org/10.1016/j.heliyon.2023.e15090
3. Blue, L., Warren, K., Abdullah, H., Gibson, C., Vargas, L., O'Dell, J., Butler, K., & Traynor, P. (2022). *Who are you (I really wanna know)? Detecting audio {deepfakes} through vocal tract reconstruction*. 2691–2708. https://www.usenix.org/conference/usenixsecurity22/presentation/blue
4. Rabhi, M., Bakiras, S., & Di Pietro, R. (2024). Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Systems with Applications*, *250*, 123941. https://doi.org/10.1016/j.eswa.2024.123941
5. (N.d.). Retrieved April 16, 2025, from https://repository.kaust.edu.sa/items/8bd32fc7-db5c-4b75-a7bb-180fb899b7b2