DATA-DRIVEN MODEL TO MITIGATE AI DEEPFAKES IN VOICE CLONES

ANT APPROACH TO ANALYZE THE FAILURE OF ASVSPOOF 2021 CHALLENGE

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Systems Engineering

By
Padma Lim

November 8, 2024

Technical Team Members:
Vishnu Lakshmanan
Drake Ferri
Rhea Agarwal
Baani Kaur
Fahima Mysha

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.

ADVISORS

Ben Laugelli, Department of Engineering and Society

Gregory Gerling, Department of Systems Engineering

Introduction

The prevalence of generative Artificial Intelligence (AI) tools enable imposters to create deepfake voices to impersonate customers, bypass security controls, and commit financial fraud ("AI Scams, deep fakes, impersonations," 2024). Financial services companies must therefore protect customer assets from such malicious activities. To achieve this goal, it is crucial to understand the current state of open-source AI voice cloning tools and the qualities that yield realistic cloned voices. My technical project aims to develop a comprehensive database that investigates the ideal combination of factors that produce the most real-life, human-sounding voices from our cloned voices. The factors we choose to focus on are gender, level of background noise, type of AI tool used, volume, and length of training audio. By identifying the factors and interaction of factors that yield the most realistic voices, banks would be able to target those specific areas to strengthen their security mechanisms.

As this technical project requires the construction of a diverse network composed of technical, social, conceptual, or other resources, it is critical to understand the mechanisms behind successful network formation to support both its development and implementation. To examine such mechanisms, I will draw on the Science, Technology, and Society (STS) framework of Actor-Network Theory (ANT) to examine how social challenges contributed to the failure of ASVspoof 2021 challenge. The challenge is an international competition focused on improving the detection of spoofed and deepfake speech in Automatic Speaker Verification (ASV) systems, which use technology to verify a person's identity based on their voice. Specifically, I will investigate how public trust in data, awareness of the dangers of deepfake technology, and collaborations between stakeholders contribute to the failure of the deepfake speech detection task. It is crucial to address both technical and social aspects of the

sociotechnical challenge as focusing solely on the technical aspects can overlook important complexities in further measures to enhance the security systems. Financial institutions, such as banks, that only develop advanced technology to detect AI-generated voice fraud but not address the social challenges risk ongoing vulnerabilities and exploitation by imposters.

Because the challenge of detecting deepfakes in financial institutions is sociotechnical in nature, it requires attending to both its technical and social aspects to accomplish successfully. In what follows, I set out two related research proposals: a technical project proposal for developing a database that shows which combination of factors in cloned voices are most successful in bypassing bank security systems and an STS project proposal for examining the social factors in the ASVspoof 2021 challenge.

Technical Project Proposal

Voice cloning is a technology that uses AI to create a synthetic voice that is almost identical to that of a real human voice. This involves training AI models using audio samples of the target voice. Within the banking industry, Automatic Speaker Verification (ASV) systems have been used to verify users' identities in recent years. An ASV system processes the voice through a microphone and either accepts or rejects the claimed identity based on the genuineness of his/her voice (Mittal & Dua, 2022, p. 4). While services are faster, malicious actors are also more likely to use cloned voices to create deepfake voice prints to impersonate customers. To mitigate these risks, countermeasures have been developed. They refer to methods designed to differentiate between bonafide and spoofed speech to improve the security and robustness of ASV systems (Kassis & Hengartner, 2023, p. 1). One countermeasure is artifact detection, which identifies anomalies in the audio that indicate synthetic generation, such as those introduced by neural vocoders (Sun et al., 2023, p. 1). Another one is liveness detection, honing in on discerning "real biometric traits from forged or replayed ones" (Kuznetsov, et al., 2021, p. 2). The third is audio watermarking, which involves embedding information into audio signals to verify its source and prevent tampering (Uddin et al., 2024, p. 2).

However, there is a problem in their design. These approaches operate in isolation and do not account for the diverse range of variables that influence the realism of synthetic speech. For example, artifact detection may not detect anomalies if the training data is not comprehensive enough; AI models that mimic human speech patterns more accurately can bypass liveness detection; watermarks may not be detectable in all situations. The current state of countermeasures does not account for the real-life scenario where multiple factors are involved to create a synthetic voice.

The aim of this project is to provide a holistic, systems-based approach that integrates various factors to enhance the effectiveness of current countermeasure designs. Specifically, I propose developing a comprehensive database that identifies the combinations of factors that produce the most realistic cloned voices. Banks can then utilize this system to identify high-risk combinations, enabling their security mechanisms to focus on and strengthen defenses against them. This database will assess the effectiveness of spoofed voice in bypassing a bank's ASV system by considering the following factors: gender, the type of AI voice cloning tools, presence of background noises, length of audio, and volume that go into producing cloned voices. By providing detailed insights into these variables, stakeholders which heavily rely on ASV for verification in the banking sector will be able to enhance their security measures, making it more difficult for malicious actors to use cloned voices to bypass authentication systems. This approach improves the detection of synthetic speech, informs the development of more robust countermeasures, and helps identify potential threats.

To develop the design, our team adopts a mix of engineering knowledge, skills, methods, and processes pertinent to systems engineering. The approach consists of an experimental design, a real-life test with an actual bank, and an optimization matrix. First, we will clone our voices and conduct an experimental design with other people to see how much they can distinguish between our real and cloned voices. Through interviews, we will gather qualitative insights into why listeners perceive the voices the way they did. For the second part, we will use Voice Over IP (VoIP), a technology that enables users to make phone calls over the internet instead of the traditional phone line, to feed our cloned voices directly to the bank's customer service number and test the robustness of its ASV system. Ultimately, we will create a model that shows the exact probabilities of our cloned voices successfully passing the ASV system. Lastly, through

conducting a sensitivity analysis, we will be able to optimize the results and find answers to the following: “(x) factor at (y) level with (z) voice cloning tool creates the most realistic voice, with (p) probability of getting past the ASV.”

Most of the initial data for determining the factors of our study consists of scholarly articles on the qualities that make up a real voice, such as stress, intonation, and rhythm (“Don’t underestimate the power of your voice,” 2022). As our team progresses, we will collect and analyze both quantitative data of the probabilities of cloned voices passing the ASV system and qualitative data of the reasoning behind users identifying cloned voices as real. This provides insights into the perceptual cues that influence their decisions. Additionally, feedback from our real-world testing will offer practical insights into the effectiveness of our cloned voices in bypassing voice authentication systems. By combining these data sources, we will refine our design, validate its effectiveness, and demonstrate its practical benefits to stakeholders. This ensures that our approach is both robust and adaptable to evolving technologies.

STS Project

ASVspoof 2021 is the fourth edition of a bi-annual challenge series aimed at advancing the study of spoofing and the development of countermeasures to protect ASV systems (Yamagishi et al., 2021, p. 1). The ASVspoof 2021 failed because it had difficulty in detecting synthetically generated speech, which contributed to the high Equal Error Rate (EER) of 22.38% for the deepfake speech detection task (Yamagishi et al., 2021, p. 5). The countermeasures developed for this challenge struggled with accurately identifying deepfake speech, which means they often misclassified spoofed speech as genuine. Additionally, the gap between promising progress results and notably worse evaluation results suggested a high degree of overfitting, indicating that the models learned patterns specific to the training data but failed to generalize to new types of deepfake attacks not present in the training data (Yamagishi et al., 2021, p. 6). Some writers argue that technical limitations in detecting synthetic speech – such as the lack of robustness to variations between simulated and real acoustic environments, and the inability to generalize across different source datasets – were primarily responsible for the project's failure (Liu et al., 2023, p. 1). While this is indeed true, this perspective often overlooks the role of social factors in the project's failure, including public trust, awareness of risks of deep fake technology, and collaborations between stakeholders. The lack of trust in the authenticity of audio and video media due to deepfake technology undermines the effectiveness of any countermeasures developed. Without proper education of the implication of deepfake technology, users may not recognize the importance of robust detection systems, leading to insufficient investment and prioritization. Furthermore, poor communication among researchers, developers, policymakers, and end-users results in fragmented efforts and a lack of cohesive strategies to address the challenges posed by deepfake technology. If these factors are not taken into account,

we will not have a comprehensive understanding of why the project failed. In the case for the ASVspoof 2021 challenge, I argue that the technical factors along with social factors regarding users and stakeholders caused the project to fail. Specifically, it is the role of and interconnections among these factors—including public perception, ethical concerns, awareness and education, collaboration and communication, and the impact on vulnerable groups—that contribute to the network's failure.

My argument draws upon the STS concept of ANT , developed by sociologists of technology – Michel Callon, Bruno Latour, and John Law. ANT analyzes how network builders construct heterogeneous networks that involve human and non-human actors to accomplish a goal(Callon, 1987, p. 11). Through using his concept of power in networks and translation (Callon, 1986, p. 18), this project will reveal how different actors, such as users, developers, regulators and the technologies themselves, contribute to the failure of the countermeasure in the ASVspoof 2021 challenge. To support my argument, I will analyze evidence from these specific primary sources from the reports in the National Science Foundation, which provide information about how public perception and awareness can influence the acceptance and effectiveness of technological solutions (National Science Foundation, n.d.). Similarly, I will also use the study that highlights how public perceptions of social and cultural impacts are often undervalued in policy decisions, which can lead to public resistance and reduced effectiveness of technological implementation (Bromley-Trujillo & Poe, 2021, p. 2). By incorporating these into the larger network of the deepfakes surrounding banks, we can enhance the effectiveness of countermeasures and spread awareness of dangers regarding deepfakes in AI.

Conclusion

In conclusion, the technical project aims to tackle the sociotechnical challenge of reducing fraud caused by deepfake AI cloned voices by developing a novel, holistic, and comprehensive database that analyzes successful cloned samples. This provides stakeholders in financial institutions concrete ways to enhance security measures for detecting spoofed voices. Concurrently, the STS project utilizes the scholarly framework of ANT to address how the power dynamics of network builders and actors contributed to the failure of ASVspoof 2021 challenge. Insights from the STS research, such as the interaction of social factors in public trust, awareness, and collaboration with technical elements, will be applied to refine the technical project. By integrating these insights, I aim to bridge the gap between users and banking systems by developing a robust and effective database. This integrated approach will ensure a comprehensive solution to the sociotechnical challenge, addressing both technical intricacies and the broader conceptual complexities.

References

Bromley-Trujillo, R., & Poe, J. (2021). Public perceptions of social and cultural impacts in policy decisions. *Proceedings of the National Academy of Sciences, 118*(17), https://doi.org/10.1073/pnas.2020491118

Bullock, D., & Sánchez, R. (2022, April 13). *Don't underestimate the power of your voice*. Harvard Business Review. https://hbr.org/2022/04/dont-underestimate-the-power-of-your-voice

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. In J. Law (Ed.), *Power, action and belief: A new sociology of knowledge?* (pp. 196-223). London: Routledge.

Callon, M. (1987). Society in the making: The study of technology as a tool for sociological analysis. In W. E. Bijker, T. P. Hughes, & T. J. Pinch (Eds.), *The social construction of technological systems: New directions in the sociology and history of technology* (pp. 83-103). Cambridge, MA: MIT Press.

Kassis, A & Hengartner, U. (2023). *Breaking Security-Critical Voice Authentication* (Publication No. 10179374) [Doctoral dissertation, University of Waterloo]. IEEE Symposium on Security and Privacy.

Kuznetsov, O., Zakharov, D., Frontoni, E., Maranesi, A., & Bohucharskyi, S. (2021). Cross-database liveness detection: Insights from comparative biometric analysis. *Journal of Biometric Research, 15*(3), 123-135. https://doi.org/10.48550/arXiv.2401.16232

Linde, I. (2024, July 9). *AI scams, deep fakes, impersonations … oh my.* J.P. Morgan. https://www.jpmorgan.com/insights/fraud/fraud-protection/ai-scams-deep-fakes-impersonations-oh-my

Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A., & Lee, K. A. (2023). *ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild*. arXiv. https://arxiv.org/pdf/2210.02437v3

Mittal, A., & Dua, M. (2022). Automatic speaker verification systems and spoof detection techniques: Review and analysis. *Int J Speech Technol, 25*(1), 105–134. https://doi.org/10.1007/s10772-021-09876-2

National Science Foundation. (2022). *Science and technology: Public perceptions, awareness, and information Sources.* Science and Engineering Indicators. https://ncses.nsf.gov/pubs/nsb20244

Sun, C., Jia, S., Hou, S., & Lyu, S. (2023). AI-synthesized voice detection using neural vocoder artifacts. [Doctoral dissertation, University at Buffalo, State University of New York] arXiv.

Uddin, M. S., Ohidujjaman, Hasan, M., & Shimamura, T. (2024). Audio watermarking: A comprehensive review. *International Journal of Advanced Computer Science and Applications, 15*(5). https://doi.org/10.14569/IJACSA.2024.01505141

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*. arXiv. https://arxiv.org/abs/2109.00537