Combatting Accessible Learning Models: Improving the Detection of Audio Deepfakes

(Technical Paper)

Navigating the Ethics of Accessible Audio Deepfake Technology

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Robert Jason Hudson

November 30th, 2023

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

Signature R.J. Hudson Date 11/30

Robert Jason Hudson

Approved _____ Date _____

Department of Computer Science

Approved _____ Date _____

STS Advisor: Richard D. Jacques, Ph.D., Department of Engineering & Society

Introduction

In today's digital era, advancements such as deepfake audio technologies have significantly blurred the line between reality and fabrication. This technology, once limited to high-end research labs, has become more accessible, allowing many people to create hyper-realistic synthetic media. This change, fueled by accessible sophisticated machine learning models, opens up immense potential for industries ranging from entertainment to personal virtual assistants. However, this advancement also presents a sinister side; the same tools that can entertain or assist may equally deceive, manipulate, and disrupt. This poses a prominent challenge to both technical and social systems across the world.

Recent studies, such as those by Mai et al. (2023), underscore a concerning reality: the average individual frequently cannot distinguish between genuine voices and their deepfake counterparts in major languages. This issue is not just a technical anomaly but a societal dilemma, raising crucial questions on trust, misinformation, and the very fabric of digital interactions. The problem frame is complex and incorporates two components: the technical intricacies involved in detecting synthetic audio and the broader societal repercussions stemming from the widespread use of deepfake audio.

At its core, the challenge of deepfake audio showcases a prime feature of what artificial intelligence (AI) can achieve; It is able to replicate or alter voice to an extent where it becomes audibly indistinguishable from the real thing. This achieved through subtype technologies like text-to-speech (TTS) and voice conversion, as well as deep learning architectures like Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs) that make these processes efficient and robust (Khanjani et al., 2023). The better these audios become, the harder

they are to detect, highlighting the need for continual advancements in synthetic audio detection technology to not only catch up with but anticipate the pace of deepfake advancements.

The proliferation of deepfake audio has frightening implications. The range of societal domains impacted by this technology is extensive and diverse; in the broadest sense, a fundamental element of human interaction is beginning to crumble. As distinguishing between genuine and synthetic becomes harder, a future where every digital interaction is met with skepticism isn't hard to envision. The potential weaponization of deepfake audios to create false narratives, deceive audiences, or manipulate popular opinion, especially in a political realm, calls for a broader examination of the societal impact if deepfake audio is left unchecked.

Technical Discussion

Deepfake audio technologies serve as an incredible testament to the advancements in artificial intelligence and machine learning, showcasing the ability to generate synthetic audio that is often indistinguishable from genuine human speech. The process of creating deepfake audio encompasses two primary sub-technologies: Text-to-Speech (TTS) and voice conversion. TTS systems translate text input into spoken words, undergoing several steps starting from text analysis to identify words, syllables, and phonemes, followed by prosody algorithms assigning pitch and duration to each phoneme, and eventually the synthesis of audio from the processed phonemes using a database of human speech recordings to generate a seamless audio output (M. P et al., 2023). Voice conversion technology, on the other hand, alters existing voice recordings to mimic another individual's voice by extracting features from the source audio, analyzing a target voice, and using algorithms to map the features of the source voice to the target voice, altering the audio data to match the desired output (Walczyna et al., 2023).

The effectiveness of deepfake audio generation is further amplified by the deployment of

advanced deep learning architectures such as Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs). GANs consist of two neural networks, a generator and a discriminator, with the generator creating new audio data resembling real audio, and the discriminator working to distinguish between real and generated audio. The iterative training process enhances the generator's ability to create realistic audio over time (Dash et al., 2021). Conversely, CNNs analyze audio data hierarchically, as they identify patterns at multiple levels of abstraction, making them effective in distinguishing subtle features and characteristics that define individual voices.

Training datasets form the backbone of deepfake audio generation, where collections of real audio data are utilized to train the deep learning models. The diversity and quality of these datasets play a pivotal role in the model's ability to learn a wide array of vocal characteristics; typically, the larger the dataset is, the better generated deepfake audio sounds. Post-processing steps following audio generation or conversion, including noise reduction and frequency response equalization, further refine the audio output, contributing to the enhanced realism and clarity of the generated deepfake audio (Mehrish et al., 2023).

There are many factors and methods at play concerning deepfake audio generation, and this exact sophistication poses substantial challenges for detection. The evolving generation techniques require modern detection methods to continuously adapt. A primary challenge in detection is the substantial computational resources required; for instance, processing audios through CNNs demands a significant amount of computation (Almutairi and Elgibreen, 2023). Some methodologies alleviate this by converting audios to images of audio features such as Spectrogram, MFCC, FFT, or STFT before feeding them into CNNs, aiming to reduce the computational burden. These detection methods employ various strategies, often leveraging

subtle discrepancies between real and synthetic audio to distinguish what is real and fake. However, it should be noted that as generation techniques become more refined, these discrepancies will become increasingly elusive, meaning more sophisticated detection algorithms are needed in the future.

Research behind deepfake audio detection is in a continuous state of evolution, striving to keep pace with the complexities of its audio generation counterpart. Within this field, my project would entail extending an existing project or evaluating a new method of detection that can be implemented on a more local scale to enhance detection accuracy. For instance, if Deep Convolutional Neural Networks were somehow harnessed within a localized detection framework it would provide an accessible solution to communities or organizations lacking large computational resources (Alzubaidi et al., 2021). The project would require optimizing the architecture of CNNs for lower computational overhead while still maintaining a high degree of accuracy.

This approach entails a rigorous evaluation and possibly reconfiguration of existing detection algorithms to suit a localized framework. By expanding this research, it's possible to find and develop a new method that leverages unique local data or insights to detect deepfake audio. Additionally, integrating other machine learning paradigms like reinforcement learning could further enhance the robustness and adaptability of the localized detection algorithms against deepfake technologies. A major roadblock, however, is how this would require a large collaborative effort between local stakeholders and technical experts to ensure the effectiveness and feasibility of the proposed solution.

STS Discussion

The integration of deepfake audio technology into the fabric of daily life presents a complex challenge across a broad societal spectrum. Trust, a basic pillar supporting human interactions, is now facing a significant test. The growing difficulty in distinguishing between real and synthetic audio could lead to an era where skepticism dominates both digital and physical interactions.

An erosion in overall trust has tangible real-world consequences. In a time when misinformation is already a challenge, deepfake audios have the potential to create false narratives, deceive audiences, and sway public opinion. Systems ingrained into everyone's daily lives are now at risk of being completely untrustworthy and therein unusable or avoided.

The potential use of deepfake audio technology as a political tool is concerning. Critically important democratic elections could be manipulated through fabricated audio clips that attribute fake quotations to prospective candidates. This has the potential to alter the developmental course of nations and governmental systems. But the problem extends beyond direct misinformation, as the atmosphere of doubt created by such technologies is also a threat. Even genuine audios could be discredited as deepfakes if they were to become more widespread and life-like, ushering in a society where objective truth is in perpetual jeopardy.

The rise of deepfake audio technology also prompts discussions about consent and privacy. Creating synthetic audio of individuals without their knowledge or consent crosses ethical boundaries, potentially causing mental distress and societal discord. While this technology showcases the innovation humans can achieve, its unchecked growth could also reveal how it endangers the wellbeing of humanity.

Another grave concern is the potential for financial fraud and criminal activities to use deepfake audio technology. Fraudsters could employ deepfake audio to impersonate executives or authority figures to authorize fraudulent transactions, divert funds, or manipulate stock prices. These activities not only result in financial losses but also undermine the integrity of financial markets and institutions (Atleson, 2023). Social engineering attacks should also be expected; criminals now have an easy way to impersonate trusted individuals in order to trick the average person into revealing sensitive information or performing actions that compromise their personal security. This form of impersonation takes phishing scams to a new level, making the deception more convincing and the potential for damage significantly higher. Due to the newfound accessibility of synthetic media creation, these scams will only become more and more common as time goes on.

The criminal use of deepfake audio technology extends to extortion and blackmail scenarios, as well. Malevolent actors could create synthetic audio clips to fabricate compromising situations or statements, using them to blackmail individuals or organizations. The fear of reputational damage could drive victims to comply with extortion demands, leading to financial loss and psychological distress (Atleson, 2023).

All these malicious uses of deepfake audio technology highlight the urgent need for robust detection methods and legal frameworks to mitigate the risks. It is urgent that clear legal guidelines concerning the creation and dissemination of synthetic audio are created as well as public awareness campaigns on the risks associated with deepfake audio. Both of these play a crucial role in safeguarding individuals and entities against fraud and other criminal activities.

The challenges and societal implications discussed herein bring to focus the core objective of this project: to pioneer an integrated deepfake audio detection technology at a more

localized level. This endeavor aims not only to bolster the resilience of communities against deceit but also to provide a model for how similar technological countermeasures can be effectively decentralized and made more accessible.

Research

The potential entanglement of deepfake audio technology within societal dynamics calls for a nuanced exploration to understand and address the challenges it presents. The primary research question guiding this project is: How can deepfake audio detection technology be effectively localized to make communities more resilient against the fraudulent use of synthetic audio?

To unravel this research question, a multi-step approach is needed. Initially, a comprehensive review will be conducted to garner insights into existing deepfake audio detection technologies and their limitations, especially concerning scalability and accessibility at a local level (Lyu, 2020). This will provide a foundational understanding of the current state of the art in deepfake audio detection and the gaps that this project aims to address.

Following the review, a technical analysis of existing deepfake audio detection algorithms will be performed. This will include assessing the computational requirements, accuracy, and feasibility of adapting these algorithms for localized implementation. Concurrently, a stakeholder analysis will be conducted to understand the unique needs and constraints of local communities concerning deepfake audio detection. This will involve engaging with community members, local organizations, and technical experts through communication and potential interviews.

Finally, the project finishes with a design and development phase. The goal will be to adapt or develop a deepfake audio detection algorithm that is optimized to be locally

implemented. This phase of the project will be iterative, involving prototyping, testing, and refining the algorithm based on feedback from technical experts and local stakeholders.

The project will employ both qualitative and quantitative methods to evaluate the effectiveness of the localized deepfake audio detection technology. Quantitative assessments will focus on the accuracy, speed, and computational efficiency of the algorithm, while qualitative evaluations will explore the usability and impact on community trust and security.

Conclusion

The endeavor to localize deepfake audio detection technology embodies a pivotal stride towards mitigating the perils posed by synthetic audio fraud. The technical deliverable, a localized deepfake audio detection algorithm, will serve as one countermeasure to those who attempt to use synthetic audio in order to deceive. On the societal front, a broader understanding of the societal implications of deepfake audio technology should be achieved, which helps illuminate the path to help create new policies and make the public aware. The completion of these deliverables should significantly address the challenges outlined in this document. By combining the technical ability to detect synthetic audio with the societal need for trust and authenticity, this project sets a foundation for a more resilient digital society in the face of advancing synthetic audio technologies.

References

Almutairi, Z., & Elgibreen, H. (2022). A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. Algorithms, 15(5), 155. doi: 10.3390/a15050155

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., (2021, March 31). *Review of Deep Learning: Concepts, CNN Architectures, challenges, applications, future directions - Journal of Big Data*. SpringerOpen. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8

Atleson, M. (2023, March 20). *Chatbots, deepfakes, and voice clones: Ai deception for sale*. Federal Trade Commission. https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale

Dash, A., Wang, G., & Ye, J. (2021). A review of Generative Adversarial Networks (GANs) and its applications in a wide variety of disciplines - From Medical to Remote Sensing. https://arxiv.org/pdf/2110.01442

Jasir M. P and Kannan Balakrishnan. 2022. Text-to-Speech Synthesis: Literature Review with an Emphasis on Malayalam Language. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 21, 4, Article 76 (July 2022), 56 pages. https://doi.org/10.1145/3501397

Khanjani, Z., Watson, G., & Janeja, V. P. (2023). Audio deepfakes: A survey. *Frontiers in big data*, *5*, 1001063. https://doi.org/10.3389/fdata.2022.1001063

Lyu, S. (2020). Deepfake Detection: Current Challenges and Next Steps. In 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 1-6). London, UK. doi:10.1109/ICMEW46912.2020.9105991.

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. (2023, June 3). *A review of deep learning techniques for speech processing*. Information Fusion. https://www.sciencedirect.com/science/article/abs/pii/S1566253523001859

Walczyna, T., & Piotrowski, Z. (2023). Overview of Voice Conversion Methods Based on Deep Learning. Applied Sciences, 13(5), 3100. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app13053100