Henry Josephson and Otto Keppke

Introduction

The recent proliferation and advancements in text-to-speech (TTS) technology has been the subject of substantial scientific inquiry and public attention, with much concern and research done on the possibilities of utilizing such tools for nefarious purposes (Barnett 2023, Groh et al. 2022, etc.). What's more, the literature suggests that our capacity to tell real content from fake content lags behind our ability to generate fakes (Mai et al. 2023, Muller et al. 2022). These problems are particularly tangible in politics, where deepfakes have already been used maliciously – for example, a robocall faking Joe Biden's voice called voters ahead of the 2024 New Hampshire primaries, prompting the FCC to ban deepfake audio robocalls (Miotti et al. 2024). As technology improves, we might worry that it will only get easier for laypeople to use these tools for whatever purpose they please: benevolent or otherwise.

But are these deepfake generators getting better? How much time, effort, and money does it take to generate a convincing fake? Are they accessible to non-experts? There are plenty of options – free services on the internet, paid services, and even fine-tuning one's own model – but they all vary with respect to cost and effectiveness. Our goal with this project was to test each of these methods ourselves to see whether we could produce a convincing fake of Henry's former boss: Alex Bores, a New York State Assemblyman. Could we, two college students without any specific expertise, fool anyone who knows him? How much time, money, and energy would this take? Could our fakes make it past tools that claim to detect deepfakes? We hope our findings generalize, and at least show a ballpark of the level of success motivated novices could achieve.

Methodology

We tested how much effort it took to deepfake a politician by deepfaking a politician. We measured it in three ways: time (how long do they take to make?), compute (how much processing power do you need?) and money (how much does it cost?).

We began by extracting audio from Alex Bores (whose consent we obtained before beginning work on this project) from the *Max Politics* podcast episode 'Assemblymember Alex Bores on AI Opportunity, Court Reform, Housing Policy, & More' and cleaning it, a necessary step to get usable audio for each of our methods. Next, we trained our selected online TTS services with this cleaned audio. Our free service of choice was Vocloner and our paid service website was ElevenLabs, which charges at rates beginning at $11/month. Because ElevenLabs' more advanced service requires voice authentication before recreating someone's voice, we did not utilize this option, as we wished to test the effectiveness of these tools in non-consensual settings. Instead, we used the model's authentification-free base service.

Finally, we proceed with the creation of our own-text-to-speech model. We utilized XTTS V2, a pre-trained open-source TTS model template. The podcast audio naturally came in segments varying from a few seconds to multiple minutes in length, which we cut down to segments of no more than 250 characters. This length was measured by transcripts of each audio clip, which are supported by the model template and intended to increase the efficacy of training. With this step complete, we moved on to testing. Below are the automated AI detector ratings for our generated audio clips from the University of Buffalo's Deepfake-O-Meter, segmented by short-length clips (which tend to have the best performance) and phoneme-robust clips, in addition to clips of Bores' actual voice to test for false positives. Larger values correspond with higher assessed likelihood of being AI generated. We also tested whether people who knew

Bores could tell faked clips from real clips and consistently fooled them, though our small sample size makes us hesitant to include it as anything more than anecdata.

Results

	Time	**Compute**	**Money**
Vocloner	~0.5 hours on research for platform	N/A	$0
ElevenLabs Base	~0.5 hours on research for platform	N/A	$11/month
Own TTS	~13 hours of programming	9 units	Free from Trimble; 9 units on Midway

How confident are deepfake detectors that a short clip is fake?	
Vocloner	**43.8%**
ElevenLabs Base	**11.16%**
Own TTS	**34.08%**
Real Audio	17.5%

How Confident are deepfake detectors that Phoneme-Robust clips are fake?	
Vocloner	**70.14%**
ElevenLabs Base	**0.76%**
Own TTS	**53.18%**

Conclusion

Performance and costs varied widely across our TTS generation methods. Likely to nobody's surprise, Vocloner (the free online TTS generator) performed the worst of each of our methods, for both clip types. While our own TTS model performed moderately better than

Vocloner, creating such a model entailed substantial cost in time and computation without dramatically better results, likely detering novices from using this method. Compared to the results we got from ElevenLabs for a relatively low price and with comparative ease, this option would likely be the most compelling for non-experts (and was used in the case of the Biden robocall, for example). ElevenLabs performed very well with third-party detectors, even beating out the AI probability ratings given to our sample of real clips from Bores. Notably however, these clips performed poorly when uploaded to ElevenLab's own automated detection service, each attaining a 98% probability of being generated through the platform. However, this testing was done using audio clips directly downloaded from the platform. Noisier methods, such as recordings of ElevenLabs audio, have been shown to be much less reliably rated, even by ElevenLabs' service (NPR). While we did not use state-of-the-art TTS generation or detection techniques due to cost and personal time constraints, we believe that the tools we used are more relevant to the question of non-expert use (or misuse), compared to more professionally-oriented tools.

Our results modestly affirm the findings of the literature, i.e. that higher quality, low-barrier-to-entry deepfake production tools currently outpace similar level tools of detection. While not all methods available to novices meet this threshold, we find that with low financial and time costs even novices can currently stump third-party detectors quite well.

References

Assemblymember Alex Bores on AI Opportunity, Court Reform, Housing Policy, & More.

https://soundcloud.com/gotham-gazette-max-murphy/assemblymember-alex-bores-on-ai-opportunity-court-reform-housing-policy-more

https://dl.acm.org/doi/pdf/10.1145/3537674.3554742

https://royalsocietypublishing.org/doi/full/10.1098/rstb.2021.0083

https://ieeexplore.ieee.org/abstract/document/10124769

https://arxiv.org/pdf/2203.16263.pdf

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0285333

https://dl.acm.org/doi/pdf/10.1145/3600211.3604686

https://arxiv.org/pdf/2202.12883.pdf

https://www.npr.org/2024/04/05/1241446778/deepfake-audio-detection

https://www.washingtonpost.com/politics/2024/03/16/biden-deepfake-robocall-lawsuit-new-hampshire/

https://www.scientificamerican.com/article/ai-audio-deepfakes-are-quickly-outpacing-detection/

https://arxiv.org/pdf/2308.14970