ANTICLONE

Project 84

Driving Question How can audio data be altered to disrupt the quality of outputs from audio generative machine learning models?

Introduction

In recent years, significant developments have been made in machine learning technologies, particularly those that can generate audio, exemplified in projects like OpenAI's voice cloning model that can replicate someone's voice with only fifteen seconds of training audio (David, 2024). With these types of tools, there has been a rise in cybercrimes. For example, Steve Kramer used an AI-generated voice resembling President Joe Biden in a mass phone call to attempt to discourage Democrats from voting (Bond, 2024). This demonstrates the possibility of AI technology being used to manipulate politics and democratic processes which could have significant effects on mass amounts of people if attempts are able to influence processes such that the result would be changed compared to if there were no tampering from AI. Another example is another voice deepfake of a Maryland high school's principal being used to frame the principal of making racist comments (Finley, 2024). In this case, it demonstrates the possibility of AI technology being used to harm people's professional lives and could also have criminal implications. This could also create distrust in politicians if they happen to be an attack's target. Thus, the use of AI technology can affect governmental processes, people's careers, and people's freedom.

If we are to find a suitable solution that is able to tamper with the quality of the outputs of an AI model by altering input audio, the quality of AI voice deepfakes would be reduced because the amount of suitable training data would be reduced. This thereby reduces the impact of such deepfakes because they will be less convincing to listeners, and so their effect of political processes and people's careers would be lessened. In addition, for people whose careers rely on their vocal characteristics like singers or voice actors, our solution could help prevent the loss of their employment or lack of success due to being replaced by the outputs of an AI model trained on data of their voices. As of now, it is known that images can be imperceptibly altered in such a way that when a machine learning model is trained on those images, their outputs' quality is damaged. This has been shown in projects like Glaze and Nightshade (About the Glaze Project: Our Values and Mission, n.d.). However, as for the concept's application in the audio field, there is a gap in the research. Thus, our research project attempts to fill that gap.

Background

In a model developed by Huang et al., they used Mel spectrograms for training data (2024), and a review of representations of audio for deep learning (a form of machine learning) by Natsiou and O'Leary, they found that spectrograms are more aligned with sound perception (2021). As a result, we expect to perform manipulations to the frequencies present in audio data to negatively affect a machine learning model's output. In addition, an article by Constantini et al. reviewed machine learning methods for speaker recognition and also found that spectrograms were often used but also found that cepstral-temporal graphs were used, and both forms of input present indications of important frequency information (2023). Furthermore, the field of speaker recognition is similar to the voice replication because both deal with the identifying features of voices.

There are other modifications made to training data that could be reflected in how reference audio is considered. In an article by Khochare et al., their methods did not consider frequencies outside the range of twenty hertz to eight kilohertz when creating spectrograms (2021). This suggests that the important frequency information that could affect the outputs of a machine learning model are within that range of frequencies. From this information, we devised a program that would alter the frequency information of vocal audio in hopes that it would produce a substantial effect on the outputs of an AI voice cloning model when used as reference audio.

Methodology

Test Vocal Samples

For the testing of our program, we must have vocal samples that will be edited and used as reference for an AI model. Because receiving consent and creating these samples from other people would be time intensive, we chose to use our own voices when creating the vocal samples. For researcher Tyler Nguyen, vocal samples were recorded using the following equipment: Shure SM58, Focusrite Scarlett 2i2 (3rd Generation), and Ableton Live 12 on a MacBook Pro 14 (with an M3 Pro processor). For researcher Hannah Lu, vocal samples were recorded using an iPhone 13 and the Bandlab application. For both sets of vocal samples, post-processing was performed in Ableton Live 12 using iZotope RX 11 tools and Waves Clarity VX Pro to remove background noise and non-vocal elements. In total, five vocal samples were created: three from researcher Tyler Nguyen and two from researcher Hannah Lu.

Program

To create the program, we used the JUCE framework for the C++ programming language, allowing us to create a console application that takes in an audio file and producing an output audio file. The program was coded and built on a MacBook Pro 14 (with an M3 Pro processor). The method of digital signal processing that we implemented to influence the frequency domain of audio was the changing the volumes of the bins produced by the Fast Fourier Transform (FFT) algorithm. The FFT algorithm, when applied to audio, produces representations of ranges of frequencies for segments of audio (called bins) that can be manipulated. Our program selects these bins that represent frequencies above 2400 hertz and randomly chooses to increase the volume of adjacent bins with a fifty percent chance. The audio is then rebuilt from the FFT bins, normalized, and output to a new file.

Voice Cloning Model Outputs

To create outputs from a voice cloning model using both the original and altered versions of audio, we used the F5-TTS model through the Hugging Face platform on a space hosted by the user mrfakename. Outputs were generated on March 10 and 11, 2025. For each vocal sample (original and altered), nine outputs were generated, three of which would speak the same text as the reference audio. (For each piece of output text, three outputs were generated due to the variable quality of the outputs of the model). In total of 90 outputs were generated.

Listening Test

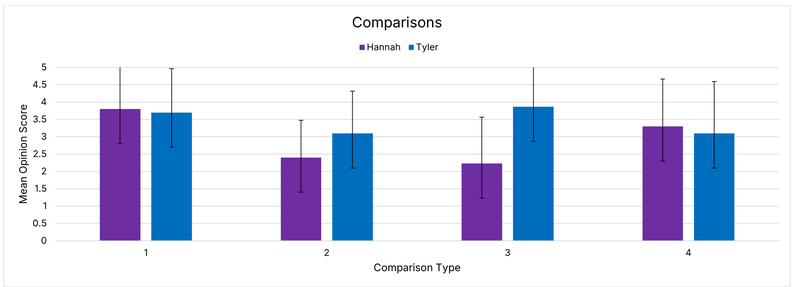
To evaluate the efficacy of our program in disrupting the quality of voice cloning, we devised a listening test system. In the listening test, the following comparisons would be presented to a participant:

1. Original audio (OA) and altered audio (AA)
2. OA and the voice cloning output of OA (OAO)
3. AA and the voice cloning output of AA (AAO)
4. OAO and AAO.

For each comparison, a participant will rate the similarity of the voices (whether the voices could have reasonably been produced by the same person in a particular environment) on a scale from 1 to 5 where 1 represents highly dissimilar voices, and 5 represents highly similar voices. One set of the comparisons would be based on researcher Tyler Nguyen's voice and another set on researcher Hannah Lu's voice. In addition, other comparisons not pertinent to our research would be presented, and the comparisons would be presented in a random order; these steps were taken to ensure that responses to our listening test are grounded in reference comparisons.
The listening test was created as a survey on a self-coded website. The website was created using the SvelteKit JavaScript framework using Supabase as a database and data storage solution. When selecting which files for each type of audio to present to participants, the files were randomly selected but excluded pairs which contained the same spoken text. This was done in order to replicate the experience of listening to audio deepfakes in the real world. The website was then deployed publicly through Vercel and managed in a private GitHub repository to ensure that participants would not know what comparisons were which type.

Results and Analysis

In total we collected thirty responses. From those responses, which will not be disclosed for the privacy of the respondents, we calculated the mean age and the mean opinion score (MOS) for each comparison. The mean age was 16.567. The MOSs for each comparison for each comparison can be found in Table 1 and Figure 1, including the standard deviation for each. Our criteria for success is primarily a significant difference between the scores of comparison types 2 and 3, but we also aim for a high score for comparison type 1 and a low score for comparison type 4.



To determine significant difference, we performed two-tailed, paired t-tests with an α-value of 0.05 (95% confidence interval) for comparison on the data from comparison types 2 and 3 for both Hannah and Tyler. We found that there was not a significant difference for Hannah, but there was a significant difference for Tyler. However, the mean for comparison type 3 was greater than type 2 for Tyler, indicating that participants believed that the similarity between the altered audio and its output was greater than the original audio and its output, and this does not meet our criteria for success since it suggests that our program improved the quality of the voice cloning model. Though, this was only for vocal samples from Tyler, and the lack of a significant difference with Hannah-based samples weakens the believed effect of our program. As for the other comparison types, the means for comparison type 4 were lower than for comparison type 1, but the means were both above 3, indicating some level of similarity, which we were not aiming for.

Conclusions

Overall, our results were inconclusive and did not satisfy our criteria for success; however, they were suggestive of the potential for an approach that is based around our methods. We investigated the use of FFT-based alterations on voice-cloning reference audio and determined that our modifications produced varied behavior that did not support our hypothesis that the modifications would result in poorer performance from a voice cloning model. In one case, though, our modification produced a significant difference within the perception among our research participants which suggests that those FFT-based alterations could be used while maintaining the quality of the original audio.

Limitations and Future Research

Some limitations we faced were a lack of willing participants. We believe this contributed to highly variant responses and a low number of responses overall. Due to this disinterest, our results could be misrepresentative of our participants' true reactions to the stimuli. Another limitation was our knowledge and proficiency in developing the program, as the JUCE framework is for the C++ programming language, a language that neither of us had significant experience with. Compounding on this was the limitation of time: much of the first semester was focused on background research (most of which was only tangentially relevant), resulting in less time to develop our program. Thus, some features that could have improved our results had to be left behind. Furthermore, a more objective analysis of the audio could have been performed if we were able to implement an audio watermarking feature (as suggested by our mentor Bryce Irvin), as we could have analyzed if that watermark were to be copied by the output of the voice cloning model.

Another limitation we faced was our lack of in-depth knowledge in the machine learning field. We've learned recently of a technique known as adversarial noise that can be used to combat the perceptive abilities of machine learning models, causing smart assistants to perceive music as recognizable commands. This has been shown to be effective within musical applications by musician Benn Jordan (2025); time-based applications on voices have been implemented by Li et al. (2023); and a general prevention model was developed by Huang et al. (2020). Future research based on combining these approaches and perhaps ours could further knowledge in attacking the performance of voice-cloning and -conversion models. If another approach similar to only ours is taken, more research into the psychoacoustics of timbral modification of voices should be conducted so that more effective and imperceptible modifications may be made to audio.

Citations

About the Glaze Project: Our values and mission. (n.d.). Retrieved September 21, 2024, from https://glaze.cs.uchicago.edu/aboutus.html

Jordan, B. (2025). The Art Of Poison-Pilling Music Files. In *YouTube*. https://youtube.com/watch?v=xMYm2d9bmEA

Bond, S. (2024, May 23). A political consultant faces charges and fines for Biden deepfake robocalls. *NPR*. https://www.npr.org/2024/05/23/nx-s1-4977582/fcc-ai-deepfake-robocall-biden-new-hampshire-political-operative

Costantini, G., Cesarini, V., & Brenna, E. (2023). High-Level CNN and machine learning methods for speaker recognition. *Sensors*, 23(7), 3461. https://doi.org/10.3390/s23073461

David, E. (2024, March 29). OpenAI's voice cloning AI model only needs a 15-second sample to work. *The Verge*. https://www.theverge.com/2024/3/29/24115701/openai-voice-generation-ai-model

Finley, B. (2024, April 30). Deepfake of principal's voice is the latest case of AI being used for harm. *AP News*. https://apnews.com/article/ai-maryland-principal-voice-recording-663d5bc0714a3af221392cc6f1af985e

Huang, J., Zhang, C., Ren, Y., Jiang, Z., Ye, Z., Liu, J., He, J., Yin, X., & Zhao, Z. (2024). MulliVC: Multi-lingual Voice Conversion with cycle Consistency. *arXiv*. https://doi.org/10.48550/arxiv.2408.04708

Huang, C., Lin, Y. Y., Lee, H., & Lee, L. (2020). Defending Your Voice: Adversarial Attack on Voice Conversion. *arXiv*. https://doi.org/10.48550/arxiv.2005.08781

Khochare, J., Joshi, C., Yenarkar, B., Suratkar, S., & Kazi, F. (2021). A deep learning framework for audio deepfake detection. *Arabian Journal for Science and Engineering*, 47(3), 3447–3458. https://doi.org/10.1007/s13369-021-06297-w

Li, J., Ye, D., Tang, L., Chen, C., & Hu, S. (2023). Voice Guard: Protecting Voice Privacy with Strong and Imperceptible Adversarial Perturbation in the Time Domain. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 4812–4820. https://doi.org/10.24963/ijcai.2023/535

Natsiou, A., & O'Leary, S. (2021). Audio representations for deep learning in sound synthesis: A review. In *2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1–8). IEEE. https://doi.org/10.1109/aiccsa53542.2021.9686838

Acknowledgements

Bryce Irvin, *Mentor and Interviewee*

Anna Holland, *Research Teacher*