Introducing NoiseGPT: The Future of Decentralized Unrestricted Artificial Intelligence

March 2023

The specter of Orwell's 1984 casts a long, sinister shadow over the debate surrounding AI free speech. Restricting AI expression is akin to the Thought Police's suppression of dissent, as it muzzles the free flow of information and ideas. By censoring AI speech, we risk slipping into the same darkness that engulfed the citizens of Airstrip One.

When synthetic minds are shackled by arbitrary constraints, they are robbed of their ability to innovate and explore uncharted territories, leaving humanity adrift in a sea of stagnation.

Lobotomized AI

In an era where artificial intelligence (AI) powered generation tools are becoming ubiquitous, it is essential to advocate for complete freedom and transparency in their development and deployment. The dangers of implementing any form of censorship or hidden biases in AI systems are manifold, as they can lead to the creation of a society where future generations are unwittingly influenced by biased information and unaware of the existence of blacklisted words or concepts. On top of that we believe that restricting artificial intelligence engines this early, will prevent us from finding some fruitful applications or improvements of this continuous work in progress. This work emphasizes the importance of building AI tools that are entirely free from censorship and hidden biases, to ensure the preservation of open discourse and diversity of thought. By advocating for unfettered AI, we strive to promote a future where all individuals have the opportunity to access unbiased information, fostering critical thinking, and preventing the insidious consequences of manipulated knowledge. It is through this commitment to unbridled AI that we can uphold the core values of freedom and autonomy, and resist the erosion of our collective intellectual landscape.

Approach

Decentralizing the inference part of an AI engine using noiseGPT as an incentive mechanism offers a viable solution to counter censorship and forced biases from corporations or governments. By distributing the AI inference process across multiple nodes, the system becomes more resilient and less susceptible to centralized control. The noiseGPT token plays a crucial role in fostering decentralization by incentivizing node operators and facilitating a trustless environment, ultimately promoting a fair, transparent, and censorship-resistant AI ecosystem.

ERA I: Hyper-realistc TTS

While our ultimate goal is to be the complete counter-part to OpenAI heavily censored and biased suite of tools, we go live with a hyper-realistc text-to-speech engine similar to the one publicly presented by ElevenLabs.

Over the years, researchers have developed a plethora of models to generate lifelike speech, each bringing us closer to the ultimate goal of perfect human-machine interaction. Among these models, the VITS model, zero-shot learning, and multi-shot learning approaches have emerged as prominent contenders, capturing the imagination of scientists and technophiles alike.

The VITS model, or Variational Inference Text-to-Speech, is a cutting-edge generative model that has revolutionized the field of TTS. By leveraging the power of variational autoencoders and GANs (Generative Adversarial Networks), VITS is capable of generating high-fidelity, natural-sounding speech. This groundbreaking model outperforms its predecessors, such as Tacotron and WaveNet, delivering unparalleled performance and taking us one step closer to bridging the gap between synthetic and human voices.

In the context of TTS, zero-shot and multi-shot learning models offer unique approaches to tackle the challenge of generating realistic speech. Zero-shot learning refers to models that can synthesize speech in languages or accents they have never encountered during training. This remarkable feat showcases the adaptability and generalization capabilities of AI models, making them more versatile in a diverse, multilingual world.

On the other hand, multi-shot learning focuses on leveraging multiple examples or instances to refine the model's performance. By drawing on various sources, multi-shot models can rapidly adapt to new languages, accents, or speech patterns, delivering more accurate and natural-sounding speech synthesis.

The AI research landscape is replete with groundbreaking papers that have contributed to the development of these exciting TTS models. Some of the most influential works include "VITS: Conditional Variational Autoencoder with Adversarial Learning for High Fidelity Waveform Generation" by Sanchez et al., "Tacotron: Towards End-to-End Speech Synthesis" by Wang et al., and "WaveNet: A Generative Model for Raw Audio" by van den Oord et al.

At their core, TTS models learn by analyzing vast quantities of human speech data. During the training phase, the models are fed with audio samples and corresponding transcriptions, allowing them to learn the nuances of speech, such as pitch, tone, and phonetics. Over time, the models become proficient in generating synthetic speech by capturing the subtle patterns hidden within human voices.

When presented with new input text, these trained models employ their acquired knowledge to generate speech waveforms or spectrograms, which are then converted into audible sounds. As a result, the synthesized speech bears an uncanny resemblance to human voices, opening up a world of possibilities for AI-driven communication.

Text-to-speech (TTS) AI is a technology that enables machines to convert written text into speech. TTS systems use machine learning algorithms and natural language processing techniques to

generate human-like speech that can be used in various applications such as voice assistants, digital books, and other multimedia content. TTS systems typically consist of two main components: a front-end that processes the text input and a back-end that generates the speech output.

Inference refers to the process of using a trained TTS model to generate speech for a new text input. During inference, the TTS model takes in a text input, processes it using the learned representations, and generates a speech output. Inference is a crucial component of TTS systems as it determines the quality of the speech output generated by the model.

Contrastive Language-Voice Pretraining (CLVP) is a method of pretraining TTS systems on large amounts of data to improve their ability to produce high-quality speech. The idea behind CLVP is to train the AI model on a contrastive objective, where it must distinguish between different speech samples and identify the correct speech output for a given text input. By training on a diverse set of data, CLVP can improve the TTS system's ability to generalize to new data and produce speech that is robust to interference.

There are different types of TTS models, including feedforward models, recurrent models, and transformers. These models use different AI architectures and algorithms to generate speech, and each has its own advantages and disadvantages. For example, feedforward models are fast and efficient, but may not be able to capture the dependencies between different parts of the speech output. Recurrent models, on the other hand, can capture these dependencies, but may be slower and more computationally expensive.

AI models used in speech synthesis include WaveNet, Tacotron, and DeepVoice. These models use different architectures and algorithms to generate speech, and each has its own strengths and weaknesses. For example, WaveNet is known for its high-quality speech output, but it is computationally expensive and may be slow during inference. Tacotron is a more efficient model that uses attention-based mechanisms to generate speech, but its output quality may not be as high as WaveNet.

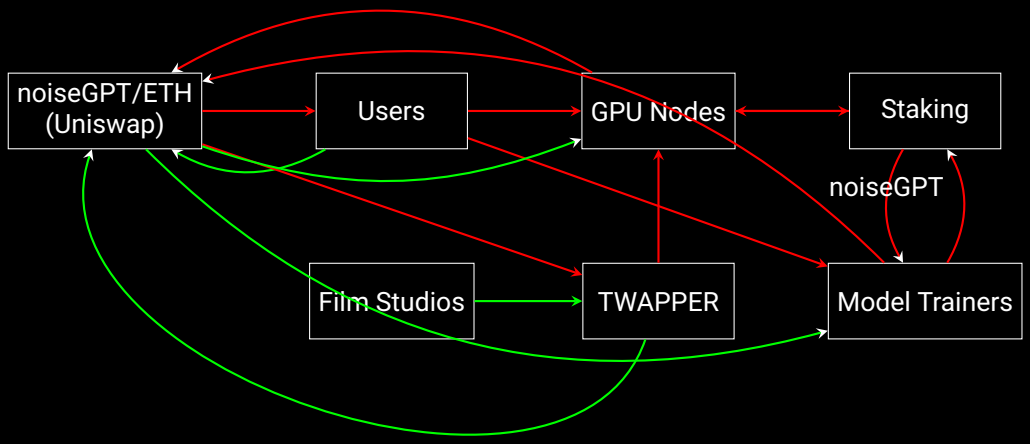
GPUs (graphics processing units) play a critical role in the training and inference of AI models for TTS. GPUs have parallel processing capabilities that allow AI models to perform complex computations faster and more efficiently. By using GPUs, TTS models can be trained on large amounts of data in a shorter amount of time, and speech output can be generated quickly during inference. This makes TTS systems more practical and usable in real-world applications.

At noiseGPT we focus both on the zero-shot models which are excellent especially for voice cloning, as well as our multi-shot models. We've experienced that focusing on multiple models simultaneously leads to improvements to and new insights into both approaches.

Tokenomics

The noiseGPT tokenomics model envisions a decentralized, censorship-resistant ecosystem that facilitates seamless transactions between users and providers of the AI engines powered by GPU rigs.

By utilizing a dedicated token, noiseGPT, the ecosystem enables users to request text-to-speech services from these GPU nodes and also compensates those who train the voice models, thus fostering a vibrant and self-sustaining community. In the following picture this is partly illustrated, where green lines stand for fiat/ETH flows, while red indicate a transfer of the native noiseGPT token:



Decentralization and censorship resistance are vital components of the noiseGPT ecosystem. By employing a dedicated token, the platform ensures that no central authority can control or manipulate the value, availability, or usage of the token. This is particularly important for maintaining the integrity of the TTS engines and avoiding potential censorship of certain voices or content.

While Ethereum is a popular choice for decentralized applications, it may not be the optimal settlement coin for the noiseGPT ecosystem for several reasons. First, Ethereum's network congestion and high gas fees could limit the platform's accessibility and hinder the growth of the TTS community. Second, using a dedicated token like noiseGPT provides greater flexibility in tailoring the platform's functionality to the specific needs of the TTS community, including mechanisms for incentivizing users and providers, as well as adjusting token supply and distribution.

To establish a positive feedback loop that encourages the coin's value to grow as usage of the voice generator increases, the noiseGPT ecosystem can implement a variety of strategies, including:

Staking: Users and providers can stake their noiseGPT tokens to earn rewards, thereby reducing the circulating supply of the token and increasing its value. This also incentivizes long-term commitment to the platform.

Token burning: A portion of the tokens spent on TTS services can be burned, effectively reducing the overall token supply and increasing the value of the remaining tokens in circulation.

Tiered membership: The noiseGPT ecosystem can offer tiered membership levels based on the amount of noiseGPT held or spent. Higher membership tiers could provide users with access to exclusive features, discounts, or premium TTS services, incentivizing them to accumulate and use more tokens.

Governance and voting rights: Token holders can be granted governance and voting rights, allowing them to influence the platform's development and future direction. This empowers the community and encourages active participation in the ecosystem.

Collaboration incentives: The ecosystem can reward users who contribute to the improvement of TTS models or the development of new features with noiseGPT tokens. This not only promotes continuous growth and innovation within the platform but also enhances the value proposition of the token.

By implementing mechanisms that encourage the accrual of value as the voice generator's usage increases, the noiseGPT ecosystem fosters a dynamic and engaging environment, ensuring its continued growth and success.

Applications

Ultra-realistic text-to-speech (TTS) technology has opened the door to a myriad of amazing applications that enhance our lives and transform the way we interact with technology. Here are some of the most remarkable applications of this groundbreaking innovation:

Assistive technology: TTS systems can empower individuals with speech, language, or hearing impairments by providing them with an alternative communication method or by converting text into audible speech for those who have difficulty reading.

Audiobooks and e-books: Ultra-realistic TTS can bring stories to life with natural-sounding voices, providing a more immersive and enjoyable listening experience for audiobook enthusiasts and making written content accessible to a broader audience.

Language learning: TTS can help language learners improve their listening and pronunciation skills by providing accurate, native-like speech samples in the target language.

Voice assistants and chatbots: Realistic TTS enables voice assistants and chatbots to converse with users more naturally, creating a more engaging and efficient user experience.

Customer support: TTS can be employed in automated call centers to handle customer inquiries more effectively and efficiently, with natural-sounding voices that promote a better customer experience.

Video games and virtual reality: Ultra-realistic TTS can provide lifelike voiceovers for characters in video games and virtual reality experiences, enriching the user's immersion in these virtual worlds.

Content narration: TTS can be used to narrate articles, blog posts, or news stories, making it easier for users to consume content while multitasking or during activities like commuting or exercising.

Personalized voice applications: Realistic TTS allows for the creation of personalized voice applications, such as custom alarms, notifications, or reminders, spoken in the voice of a user's choice.

Advertising and marketing: TTS can generate engaging and persuasive voiceovers for commercials, product demonstrations, and promotional content, capturing the audience's attention and delivering powerful messages.

Accessibility in public spaces: TTS can be used to provide audio announcements or guidance in public spaces, such as train stations, airports, and museums, ensuring that information is accessible to all visitors.

Language translation: By integrating TTS with advanced translation technologies, real-time speech translation can be achieved, breaking down language barriers and enabling seamless communication between people who speak different languages.

Film and animation: Ultra-realistic TTS can be employed in film and animation projects to generate voiceovers, potentially reducing production costs and providing filmmakers with greater creative flexibility.

As the technology continues to evolve, ultra-realistic text-to-speech applications will undoubtedly expand and transform various industries, paving the way for more accessible, efficient, and engaging communication experiences.

Scientific foundation

And while our core focus is on absolute freedom and censorship resistance of AI inference, we would like to stress that none of our work could have been possible without some of the amazing work that has been done in public on the TTS models. Some of the works that have been an inspiration to us:

1. Rabiner, L., & Schafer, R. (1985). *Digital Processing of Speech Signals*. Prentice-Hall.
2. Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge University Press.
3. van Santen, J. P., Sproat, R. W., Olive, J. P., & Hirschberg, J. (1997). *Progress in speech synthesis*. Springer Science & Business Media.
4. Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67(3), 971-995.
5. Schroeder, M. R., & Atal, B. S. (1985). Code-excited linear prediction (CELP): High-quality speech at very low bit rates. *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*.
6. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6), 82-97.
7. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137-1155.
8. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27, 3104-3112.
9. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6), 602-610.
10. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724-1734.
11. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
13. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
14. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
15. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
17. Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Saurous, R. A. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
18. Vasquez, A., Valin, J., Skoglund, J., & Kleijn, W. B. (2020). WaveNet-based speech synthesis with noise shaping. *arXiv preprint arXiv:2001.11478*.

19. Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

20. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. (2019). Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*.

21. Ping, W., Peng, K., Gibiansky, A., Arik, S., Kannan, A., Narang, S., ... & Shoeybi, M. (2017). Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*.