Recent Advances in Audio Generation: From Sequence Models to DDPMs

Michael Ohagwu
Department of EECS(Undergraduate)
University of Ottawa
Ottawa, Canada
Mohag018@uottawa.ca

Macdonald Zaheer
Department of EECS (Undergraduate)
University of Ottawa
Ottawa, Canada
gmacd083@uottawa.ca

***Abstract*—This paper, Advances in Audio Generation: From Sequence Models to DDPMs, is a high-level minimally technical review of the recent advances in the realm of arbitrary audio generation systems in the space of machine learning. We will explore the underlying mechanisms allowing for such high-fidelity few-shot outputs, as well as look into the potential implications of such a system if used by wrongdoers in acts of malfeasance. We will also explore solutions to counteract any such malfeasant acts in order to maximally reap the benefits of this new technology while simultaneously minimizing damage.**

***Index Terms*—TTS, DDPM, GAN**

I. Introduction

The past decade has seen immense gains, as well as dramatic breakthroughs in the space of arbitrary audio synthesis conditioned on text-based inputs. More specifically, text-to-speech systems, as well as text-to-music have seen commendable improvements, primarily powered by algorithmic advances in the space of neural networks and end-to-end probabilistic modeling. In this paper, Advances in Audio Generation: From Sequence Models to DDPMs, we provide a survey of the underlying mechanisms of such systems, as well as the likely negative impacts of unabated use of these models, and solutions to minimize such instances. Generally speaking, TTS systems, as of recent times, operate under two fairly distinct paradigms in terms of how they are architected; namely, the cascaded text-to-speech (TTS) systems [Shen et al., 2018, Ren et al., 2019, Li et al., 2019], which primarily leverages pipelines having acoustic models and a vocoder using a mel spectrogram as an intermediate representation to be learned, that would map to an ideal waveform corresponding to the desired output. We also now have text-to-speech (TTS) systems that are based on sequence modeling of discretized audio codec codes based on phenome and acoustic code prompts that map onto the target content and the input speakers' voice [Brown et al., 2020]. The emergence of Denoising Diffusion probabilistic models (DDPMs) has evoked what some might call a renaissance in the space of generative machine learning. Initially, the positive effects of DDPMs have been thought to be constrained to the realm of generative sequence modeling and image generation, but more recent works like Audio-diffusion have shown great improvements in the output fidelity of their DDPM UNET based upsampler and downsampler block [4]. In subsequent sections of this paper, we investigate, on a deeper - more technical plane, how these paradigms of text-to-speech (TTS) systems function, as well as Audio-diffusion, and how best to leverage their fairly distinct architectures in order to provide failsafes and contingencies in order to minimize the likelihood of such systems being used for malfeasant acts.

Identify applicable funding agency here. If none, delete this.

II. System overview

TortoiseTTS, from an architectural perspective, can be conceived as five separately trained neural networks that have been pipe-lined together to achieve high-fidelity audio few-shot outputs conditioned on textual input. It includes, as its sub-networks an autoregressive decoder based on GPT(Generative Pre-Trained Transformer), CLVP and CVVP(contrastive voice-voice pre-training) embedding models, a DDPM-based decoder, and, finally, a Univnet-based neural vocoder. These sub-networks in conjunction allow for relatively high-fidelity speech audio outputs based on few-shot inputs.

In multiple ways, the open-source text-to-speech system, AudioDiffusion is similar in architectural design to TortoiseTTS, the primary caveat though is in the diffusion-based upsampler applied to the waveform, as well as the UNet-based vocoder for decoding mel-spectrograms into high-fidelity music output [4].

Now, we arrive at the final sort of text-to-speech architecture - the Large language model(LLM) based systems. These are the latest architectural paradigm being explored as viable methods for high-fidelity speech production given minimal input examples (few/zero-shot systems). The main caveat with such Language-model based text-to-to speech systems is in the intermediate representation used to train the sub-neural networks involved; namely, the use of an audio codec as that intermediate representation [Wang et al., 2020], as opposed to a mel-spectrogram as is the case in cascaded text-to-speech-systems, such as TortoiseTTS.

A. Cascaded text-to-speech (TTS) systems for zero-shot TTS

TortoiseTTS has as one of its underlying network modules, a decoder-only generative pre-trained transformer (GPT) network. This submodule is used primarily as an autoregressive

decoder and is the crux of much of the functions of the system. The system is based on a generative model that uses autoregressive decoding to produce highly-compressed audio data from text inputs and reference clips. The components of the Tortoise system work together to generate natural-sounding, high-fidelity speech outputs that most closely match the input text and reference audio it's being conditioned on.

The autoregressive decoder is the core component of the Tortoise system. It takes in text inputs and reference clips and generates latents and corresponding token codes that represent highly-compressed audio data. The latents are then used by the diffusion decoder to produce a MEL spectrogram that represents the speech output.

To generate natural-sounding speech, the Tortoise system uses a nucleus sampling decoding strategy. This means that the system generates multiple "candidate" latents for each input text and reference clip. The system then uses the CLVP and CVVP models to select the best candidate.

The CLVP model produces a similarity score between the input text and each candidate code sequence. The CVVP model produces a similarity score between the reference clips and each candidate. The two similarity scores are then combined with a weighting provided by the Tortoise user. This allows the system to choose the candidate with the highest total similarity to proceed to the next step.

Once the candidate has been selected, the diffusion decoder consumes the autoregressive latents and the reference clips to produce a MEL spectrogram representing the speech output. Finally, a Univnet vocoder is used to transform the MEL spectrogram into actual waveform data that can be played back as speech.

Overall, the Tortoise system is a highly sophisticated text-to-speech system that uses advanced machine learning techniques to produce natural-sounding, high-fidelity speech output that would most closely match the input text and reference audio. The system's autoregressive decoding, nucleus sampling, CLVP, CVVP, and diffusion decoder components all work together seamlessly as subnetworks or cascaded networks to create a highly effective and efficient speech synthesis pipeline.

B. Cascaded text-to-music(TTM) systems for zero-shot TTA

This section focuses on the underlying mechanism of Audio Diffusion based systems. Text-to-audio (TTA) systems are a technology that has recently gained attention due to their seemingly remarkable ability to synthesize high-fidelity general audio output based on text descriptions as input. Although, it must be noted that, prior studies in the space of TTA have reported high-frequency instances of limited generation quality despite the high computational costs required. This is where the novel text-to-audio (TTA) system named AudioLDM comes in.

AudioLDM is a more recent - novel text-to-audio (TTA) system built on a latent space, which means that it learns continuous audio representations from contrastive language-audio pretraining (CLAP) latents. By using pre-trained CLAP

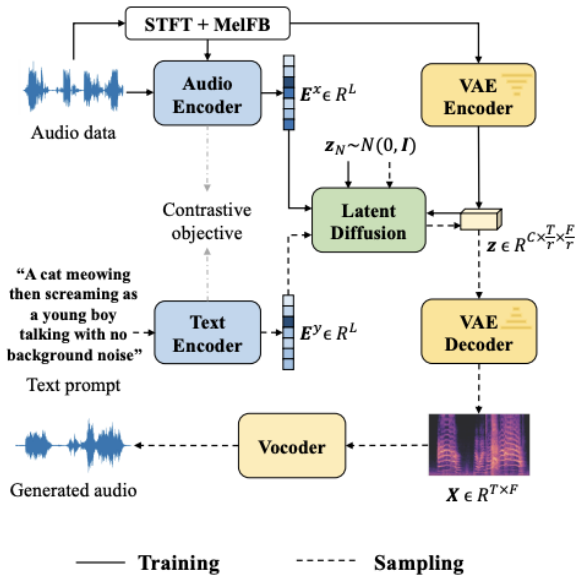

Fig. 1. An Overview of AudioLDM design for the text-to-audio generation. During training, latent diffusion models (LDMs) are conditioned on audio embedding and trained in a continuous space learned by VAE. The sampling process uses text embedding as the condition. Adapted from [2]

models, we can train LDMs (latent discriminative models) with audio embedding and text embedding as a condition during sampling. This allows AudioLDM to learn the latent representations of audio signals and their compositions without modeling explicitly the cross-modal relationship, which makes it advantageous in both generation quality and computational efficiency.

In addition, AudioLDM is trained on AudioCaps with a single GPU, yet it achieves state-of-the-art TTA performance measured by both objective and subjective metrics such as the Frechet distance. What's more, AudioLDM is the first TTA system that enables various text-guided audio manipulations, such as style transfer, in a zero-shot fashion.

Overall, the proposed AudioLDM system presents a significant improvement over previous TTA systems, as it achieves both high-fidelity generation quality and computational efficiency. Furthermore, its ability to perform text-guided audio manipulations in a zero-shot fashion makes it a highly versatile tool that could be used in various applications, such as virtual assistants, audiobook production, and speech therapy, among others.

C. LLM based systems

For this section on language model-based zero-shot text-to-speech system, we will use as our case of study, VALL-E, a zero-shot high-fidelity text-to-speech system that is predicated on using audio codec tokens for intermediate representations, as opposed to mel-spectrograms in cascaded architectures. We can properly conceive VALL-E, then as a neural codec language model such that the neural network tokenizes the

input speech and proceeds to use intermediary networks to use those output tokens to build waveforms that correspond to the voice of the speaker, including keeping the speaker's timbre and emotional tone.

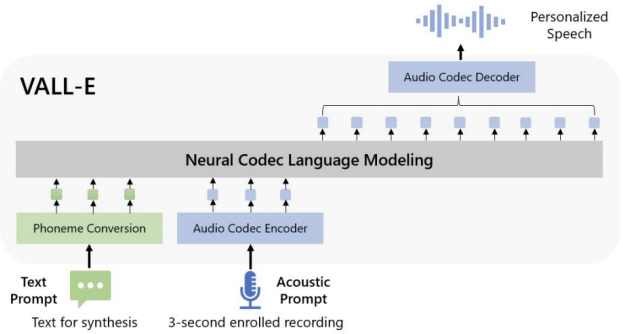

Fig. 2. The overview of VALL-E. Unlike the previous pipeline (e.g., phoneme → mel-spectrogram → waveform), the pipeline of VALL-E is phoneme → discrete code → waveform. VALL-E generates the discrete audio codec codes based on phoneme and acoustic code prompts, corresponding to the target content and the speaker's voice. VALL-E directly enables various speech synthesis applications, such as zero-shot TTS, speech editing, and content creation combined with other generative AI models like GPT-3. Adapted from [1]

Some more stuff about neural-codec based systems.

D. Comparison between Cascaded text-to-speech systems and language model text-to-speech systems

At the time of writing, neural-codec based language model text-to-speech systems seem to surpass all other text-to-speech systems on the general benchmarks corpus' such as LibriSpeech and VCTK.

	Current Systems	**VALL-E**
Intermediate representation	mel spectrogram	audio codec code
Objective function	continuous signal regression	language model
Training data	$\leq$ 600 hours	60K hours
In-context learning	✗	✓

Fig. 3. A comparison between VALL-E and current cascaded TTS systems. Adapted from [1]

III. Ethical Implications Overview

Such technologies as the ones explicated above pose a latent risk given the ethical implication accompanying their proliferation and unregulated use by both malicious individuals, as well as unethical corporations or autocratic governments. The negative ethical implications of the proliferation of audio-generative models could metastasize as - corporate espionage, use by autocratic governments to exert control, cases of fraud, and as a means to incite violence of any form. The following sections will be a more in-depth survey of each of these ethical implications, as well as possible solutions to curb any such negative outcomes that could occur as a result of proliferation and unregulated access to such powerful systems.

A. Corporate Espionage

As an initial case study for the potential negative use cases of such audio-generative systems, corporate espionage can be conceived as a low-hanging fruit with respect to the myriad of potential malfeasant acts that can be carried out in a scenario of unfettered access to such systems.

To paint an exemplifying illustration, take, for example, the scenario of where a malfeasant actor(s), either within the corporation or an external actor(s), could employ such audio-generative systems to create convincing, high-fidelity audio recordings of both confidential conversations and/or meetings held within the said corporation, which would ultimately be used for nefarious purposes.

This could ultimately include acts such as leaking sensitive information, blackmailing employees, or even as a means to manipulate stock prices. Furthermore, audio-generative systems could also be used to create fake audio evidence, in an attempt to frame certain individuals for supposed illegal activities or to impersonate high-level executives in order to gain access to even more sensitive information.

B. Autocratic Governments

A case for utilizing these new audio-generative systems by autocratic governments, in an effort to more effectively exert control on the populace via propaganda and intimidation. To summarize the primary premise, in the hands of autocratic governments, audio-generative systems can be used to perpetuate propaganda, disinformation campaigns, and psychological warfare, with devastating consequences for their citizenry and beyond.

To explicate the negative effects of autocratic governments utilizing these systems, they can:

- Dissemination of False Information: One of the most obvious negative effects of audio-generative systems being utilized by autocratic governments is the dissemination of false information. These systems can be used to create realistic audio recordings of people saying things they never actually said, or to generate sound effects that make it seem like events occurred that never actually did. This can be used to manipulate public opinion, sow discord, and undermine democratic institutions.
- Amplifying Hate Speech and Incitement to Violence: Autocratic governments can use audio-generative systems to create hateful or inflammatory messages that are designed to incite violence and discord. These messages can be amplified through social media and other digital platforms, leading to an increase in hate crimes and other forms of violence.
- Undermining Free Speech: Audio-generative systems can be used to create deepfake audio recordings that are designed to discredit journalists, activists, or other individuals who speak out against autocratic regimes. These deepfakes can be used to silence dissent and undermine free speech, making it more difficult for people to speak out against injustices.

- Creating Chaos: Audio-generative systems can also be used to create chaos and confusion. For example, autocratic governments could use these systems to create fake emergency broadcasts or sound effects that simulate explosions, gunfire, or other violent events. This could lead to mass panic and disruption, making it easier for governments to consolidate power and control their populations.
- Eroding Trust: Lastly, the use of audio-generative systems by autocratic governments erodes trust in institutions and democratic processes. When people cannot trust what they see or hear, they are more likely to become disillusioned with democratic institutions and more susceptible to authoritarianism. This could lead to the erosion of democracy and human rights, both within autocratic regimes and in other countries around the world.

So, we see that in the case of autocratic governments, these systems could have very negative impacts on human rights and democratic institutions. As these systems continue to evolve, becoming more sophisticated, we would do well to develop sufficient safeguards to prevent their misuse by such authoritarian regimes.

C. Fraud

As we are aware, audio-generative systems have the perturbing potential to create highly realistic - convincing fake audio recordings of individuals saying things that have not actually been said. Although there are multiple positive applications of this technology, such as in the entertainment industry, or more creatively, to implement high-fidelity virtual assistants. It still poses a major risk for cases of fraud.

Of the more significant negative cases of audio-generative systems being used for malfeasant acts is the case of it being utilized for fraud, which can ultimately cause serious harm to individuals and organizations. Take, for example, a malfeasant actor, in this case, a fraudster, employing the use of audio-generative technology to create fake audio recordings of a CEO or high-ranking executive giving instructions to transfer large sums of money to a given fraudulent account. This would ultimately result in significant financial losses for the organization, as well as reputational damage.

Another negative effect of audio-generative technology being used for fraud is that it can be challenging to detect. Traditional methods of verifying the authenticity of audio recordings, such as voice recognition software or visual cues, may not be effective when dealing with highly sophisticated audio-generative technology. This can make it challenging for individuals and organizations to protect themselves against fraud, as they may not even realize that they are dealing with a fake recording until it is too late.

Finally, the use of audio-generative technology for fraud can undermine public trust in the authenticity of audio recordings and other forms of evidence. If individuals and organizations cannot be confident that the audio recordings they are dealing with are authentic, it can erode trust in the justice system and other important institutions. This could have significant implications for society as a whole, as it could make it more difficult to prosecute criminals and uphold the rule of law.

Overall, the negative effects of audio-generative systems being utilized for cases of fraud are significant and far-reaching. It is essential that individuals and organizations take steps to protect themselves against this risk, such as by implementing robust authentication and verification procedures for audio recordings and other forms of evidence. Additionally, continued research and development in the field of deep fake detection will be crucial to ensuring that this technology is not used for fraudulent purposes.

D. Inciting Violence

One of the most significant negative effects of audio-generative systems being utilized for cases of inciting violence is that they can contribute to the spread of hate speech and extremist ideologies. By generating audio content that promotes violence, hatred, and discrimination against certain groups or individuals, these systems can amplify dangerous ideas and exacerbate tensions in society.

Another negative effect is that these systems can be used to create fake audio content that is indistinguishable from real recordings. This can make it difficult to determine the authenticity of audio content, leading to confusion, misinformation, and a loss of trust in media and information sources.

Furthermore, the use of audio-generative systems for inciting violence can have a chilling effect on free speech and democratic values. If individuals and groups fear that their speech and ideas may be targeted by these systems, they may self-censor or refrain from expressing themselves, leading to a suppression of diverse viewpoints and ideas.

Finally, the use of audio-generative systems for inciting violence can also have legal consequences. In many countries, hate speech and incitement to violence are considered serious crimes, and individuals or groups who use these systems to spread such content may face prosecution and punishment.

In conclusion, the negative effects of audio-generative systems being utilized for cases of inciting violence are numerous and far-reaching. It is crucial to regulate and monitor the use of these systems to prevent their misuse and ensure that they are only used for ethical and legitimate purposes.

IV. Risk mitigation systems: Discriminator

Given the myriad risk factors associated with the emergence of these powerful audio-generation techniques, we would do well to create appropriate risk mitigation techniques.

The most prominent solution to mitigate risks accompanying the proliferation of such systems would be to implement a discriminator as explicated in the canonical Generative Adversarial Network (GAN) architecture[6]. Take, for example, the two subnetworks of a canonical GAN, ie, the generator network, and the discriminator network. The discriminator system is trained against the outputs of the generator under the paradigm of a binary classifier. So, we could simply employ a similar method to effectively discriminate against the outputs of the audio waveforms generated by such audio-generative

systems. Training a discriminator network on generated audio waveforms would only likely be limited by available data and compute resources.

V. Conclusion

To conclude, the emergence of these new technologies is more than definitely going to bring about another great Khunian divergence in terms of how we orient ourselves across a wide range of social and economic institutions. We would do well to have failsafes in place in order to minimize the likelihood of negative outcomes and reap the mostly positive rewards for the betterment of humanity and society at large.

References

[1] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural Codec Language Models are Zero-Shot Text To Speech synthesizers," arXiv:2301.02111 [cs.CL], 2023. In progress.

[2] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models," arXiv:2301.12503 [cs.SD], 2023. In progress.

[3] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform Generation," arXiv:2106.07889 [eess.AS], 2021.

[4] F. Schneider, "Archisound: Audio generation with diffusion," Master's Thesis, ETH Zurich, Switzerland, Jan. 2023. Unpublished.

[5] B. James, "Spending compute for high-quality TTS", 2O22. Unpublished.

[6] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative Adversarial Networks. arXiv preprint arXiv:1406.2661 [stat.ML]. Published