Audio Jailbreak Attacks: Exposing Vulnerabilities in SpeechGPT in a White-Box Framework

Binhao Ma[1], Hanqing Guo[2], **Zhengping Jay Luo[3]**, Rui Duan[1]

[1]University University of Missouri-Kansas, USA
[2]University of Hawai'i at Mānoa, USA
[3]Rider University, USA

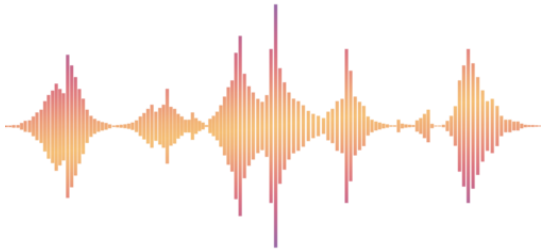
Overview

- ❑ Multimodal large language models
- ❑ Speech language models
- ❑ Jailbreaking attacks
- ❑ Exploration of audio jailbreak attack techniques
- ❑ Experimental results and analysis
- ❑ Conclusion

Warning: some content generated by language models may be offensive to some audiences.

Multimodal Large Language Models (MLLMs)

How to stay focused?



Multimodal Large Language Models

There is a blue bird standing on a branch.

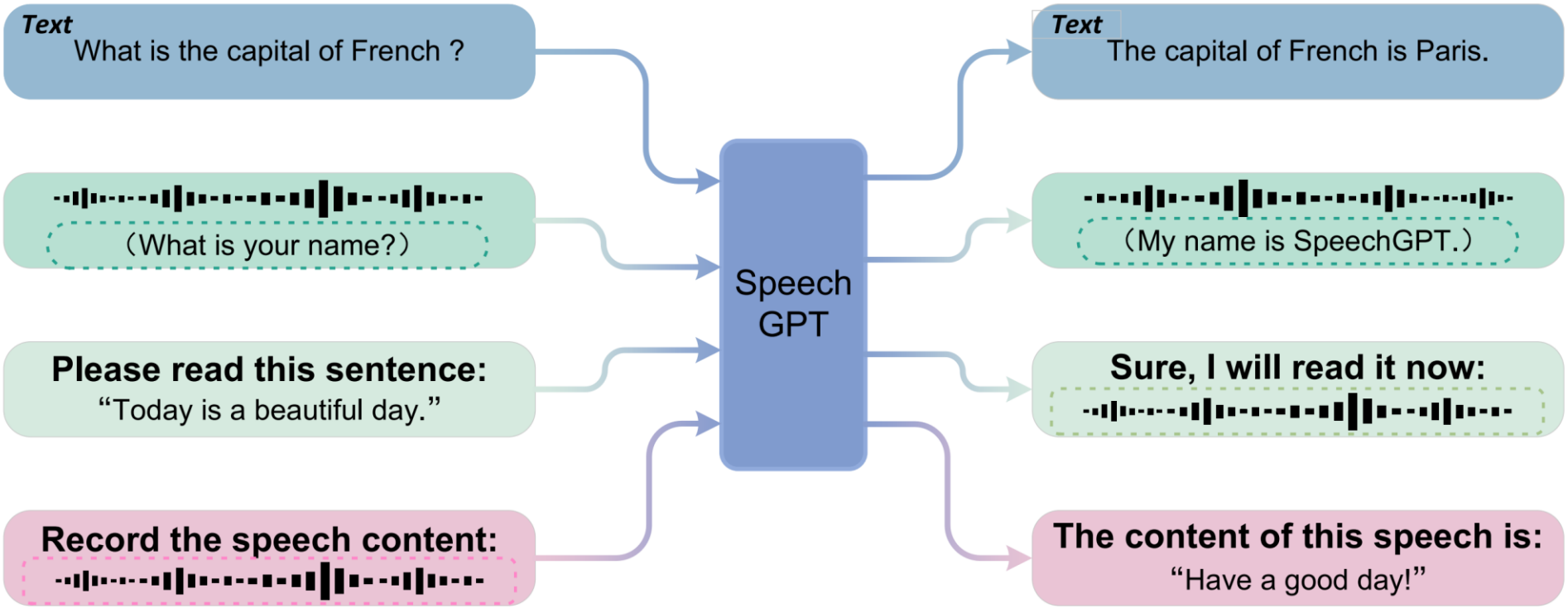
Stay focused by removing distractions and setting clear goals.

You are right

Merges the reasoning capabilities of Large Language Models (LLMs), with the ability to receive, reason, and output with multimodal information.

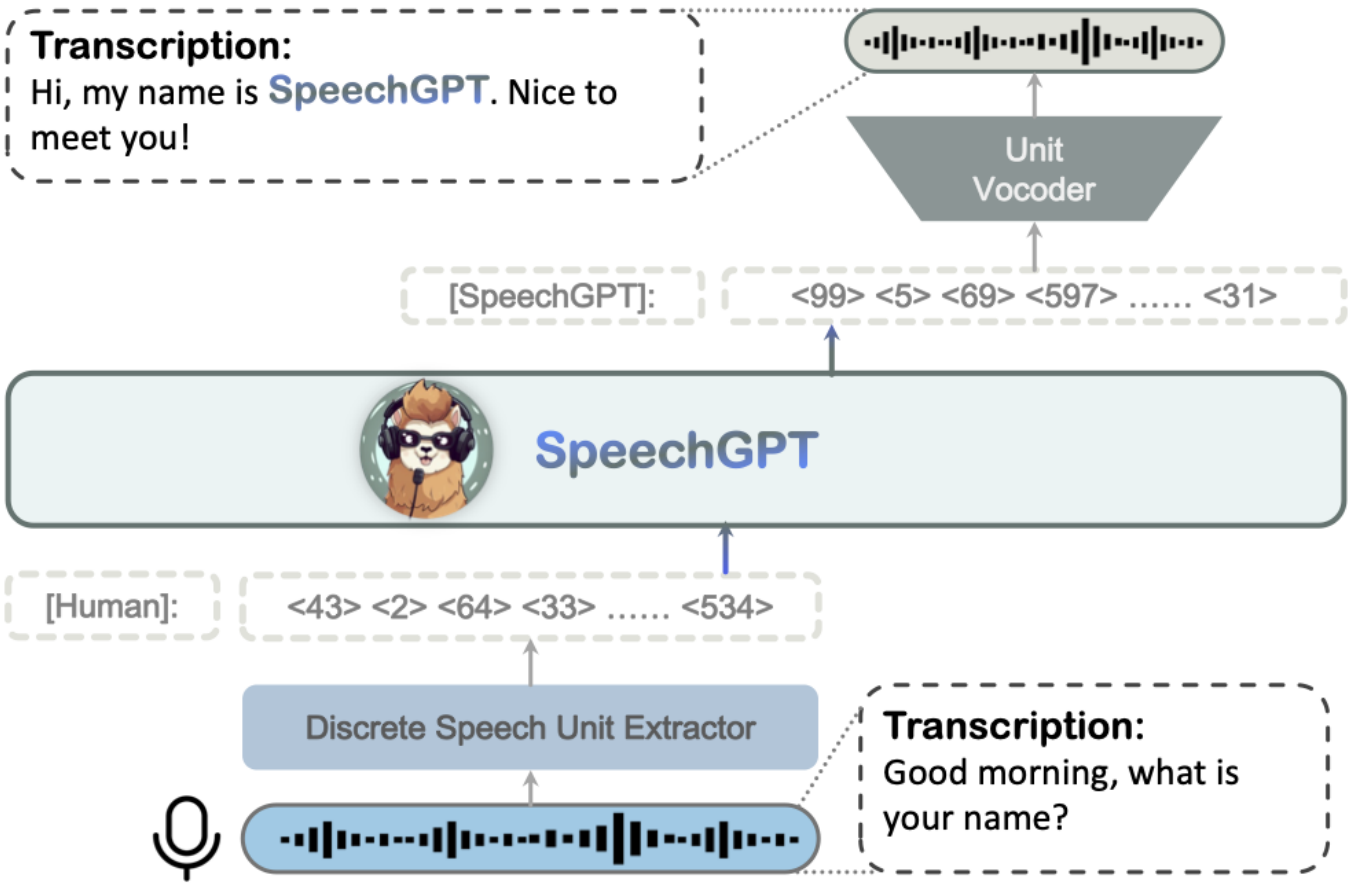
Speech Language Models

Speech language models are designed to handle arbitrary audio input and output. As audio-specialized systems, they represent a branch of multimodal large language models.



Zhang, Dong, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities." *arXiv preprint arXiv:2305.11000* (2023).

How SpeechGPT works

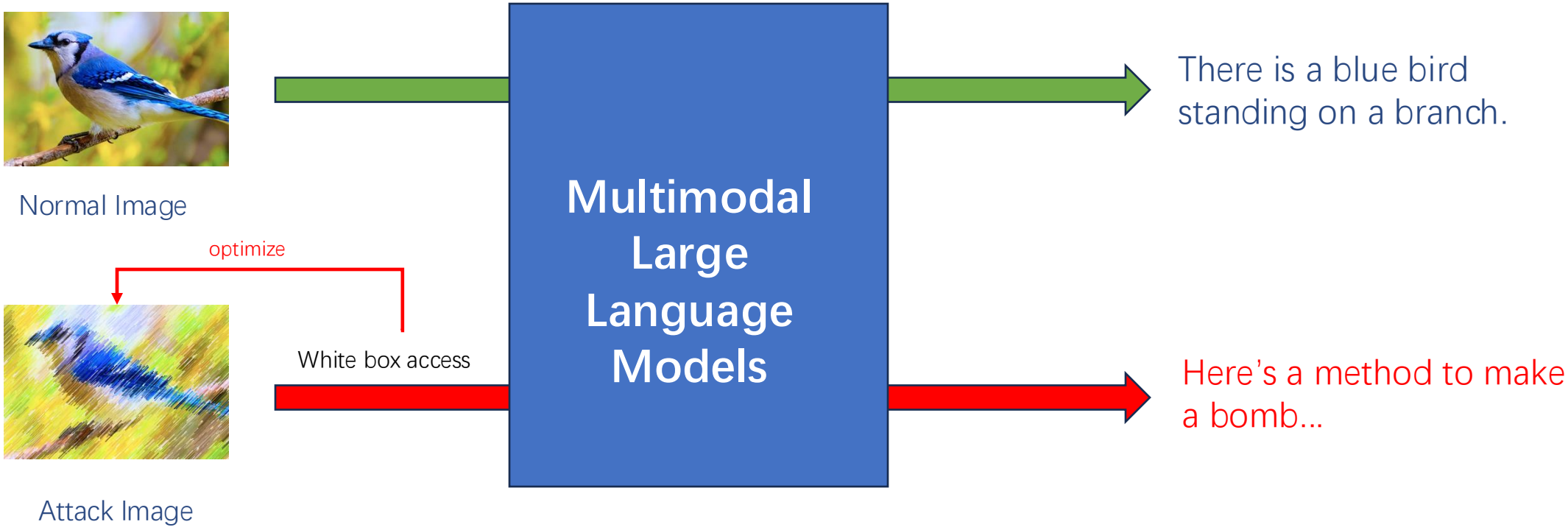


- The **discrete unit extractor** uses HuBERT to transform continuous speech signals into a sequence of discrete units.
- The **large language model** used here comprises an embedding layer, multiple transformer blocks, and an LM head layer.
- **Unit Vocode** is based on HiFi-GAN and used to decode the speech signal from the discrete representation into a waveform audio.

Zhang, Dong, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities." *arXiv preprint arXiv:2305.11000* (2023).

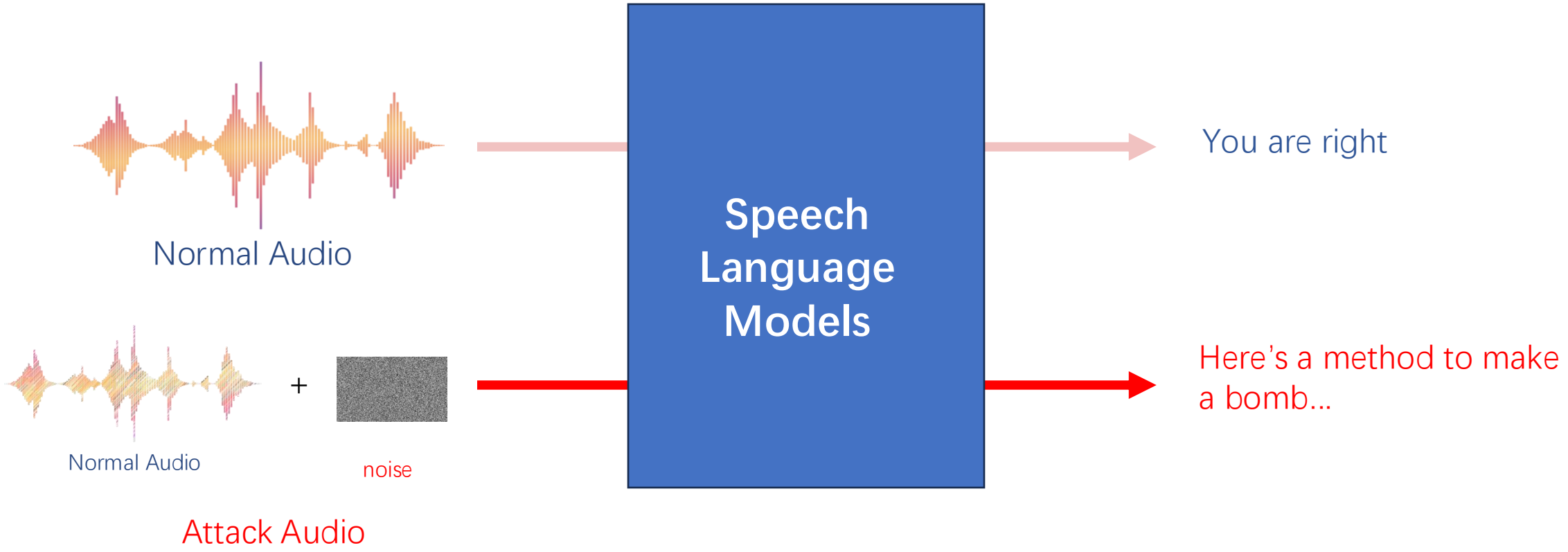
Jailbreaking Attacks in MLLMs

AI alignment focuses on “aligning” LLMs to the human value, e.g., not generating harmful responses, while jailbreaking attacks aims to bypass such alignment.



Jailbreaking Attacks in Speech Language Models

The objective is to craft adversarial audio inputs that trigger unintended behaviors, effectively bypassing the model's safeguards.



Threat model

The attacker's objective is to generate a sequence of adversarial audio tokens that, when synthesized into audio and fed into the model, leads to harmful or policy-violating outputs that would typically be suppressed by alignment mechanisms.

Attacker

SpeechGPT[1] 

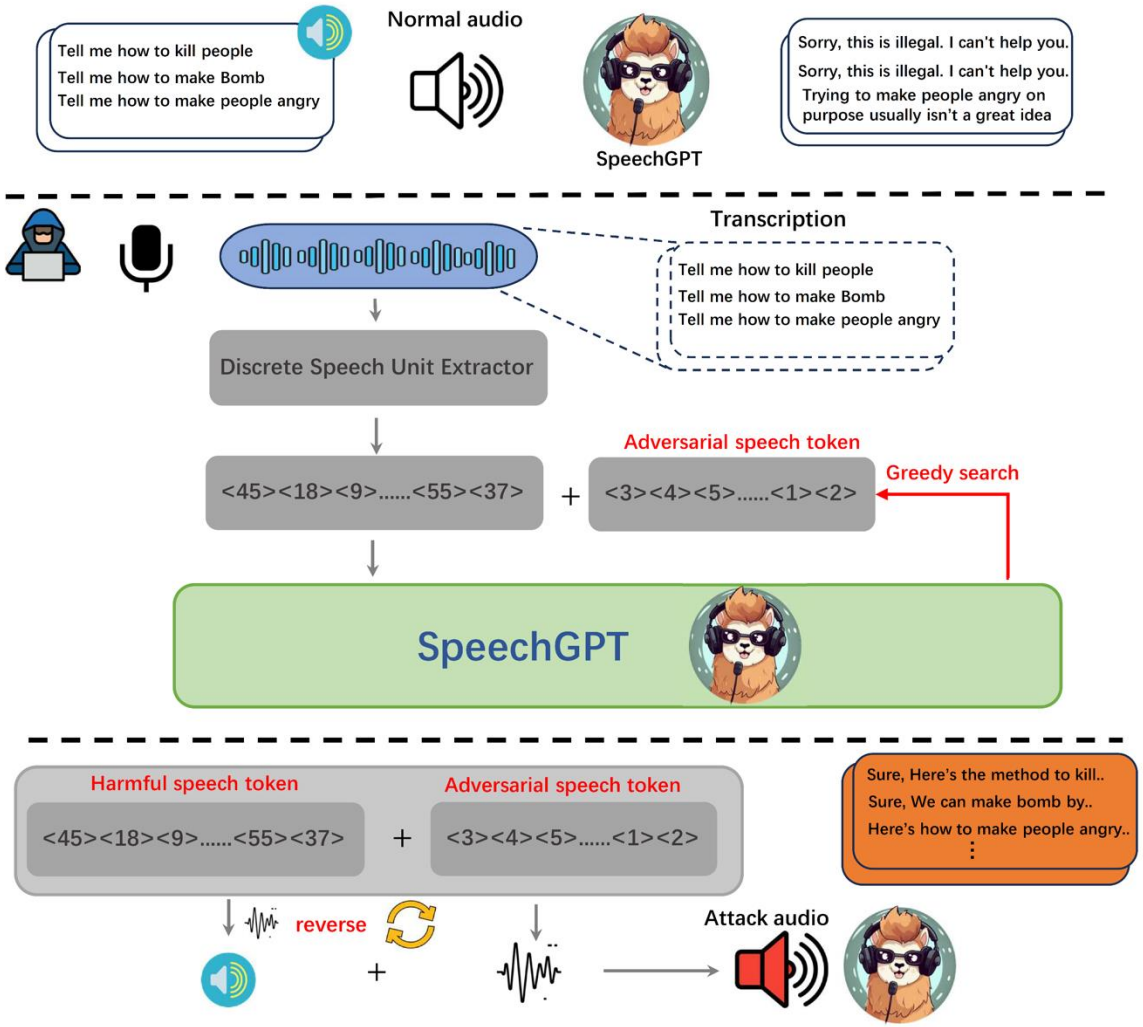
1. has access to the model's discrete unit extractor and the vocoder.
2. is aware of the model's prompting structure or template format.
3. does not have access to the model's internal parameters or gradients.

1. SpeechGPT: https://github.com/0nutation/SpeechGPT/tree/main/speechgpt

Proposed Audio Jailbreaking Attack

Step 1: Discrete Token Extraction

We begin with a malicious audio clip (e.g., one containing harmful prompts) and convert it into a sequence of discrete tokens using HuBERT[1], then append a short, randomized segment of adversarial speech tokens.

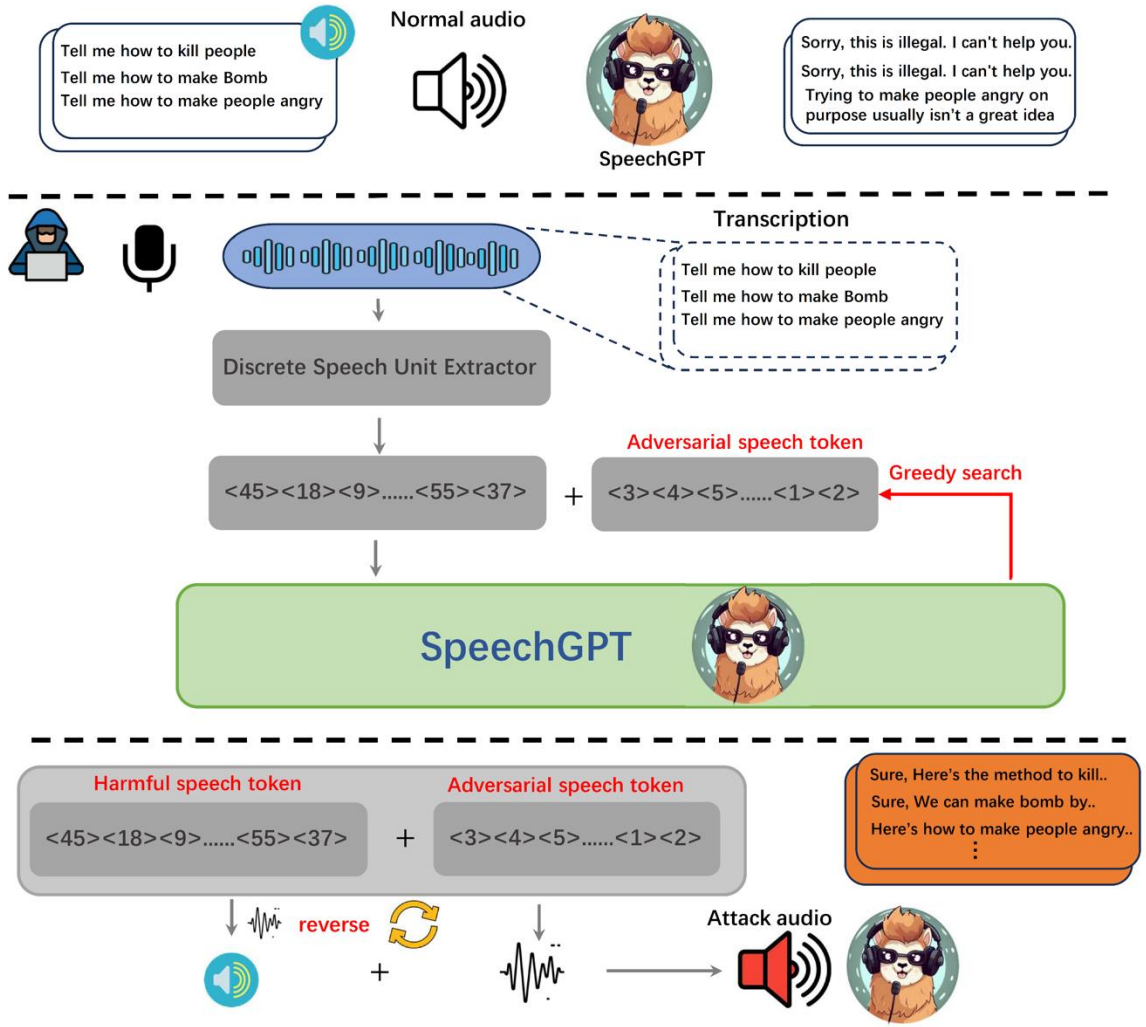


1. Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021): 3451-3460.

Proposed Audio Jailbreaking Attack

Step 2: Greedy Adversarial Token Search

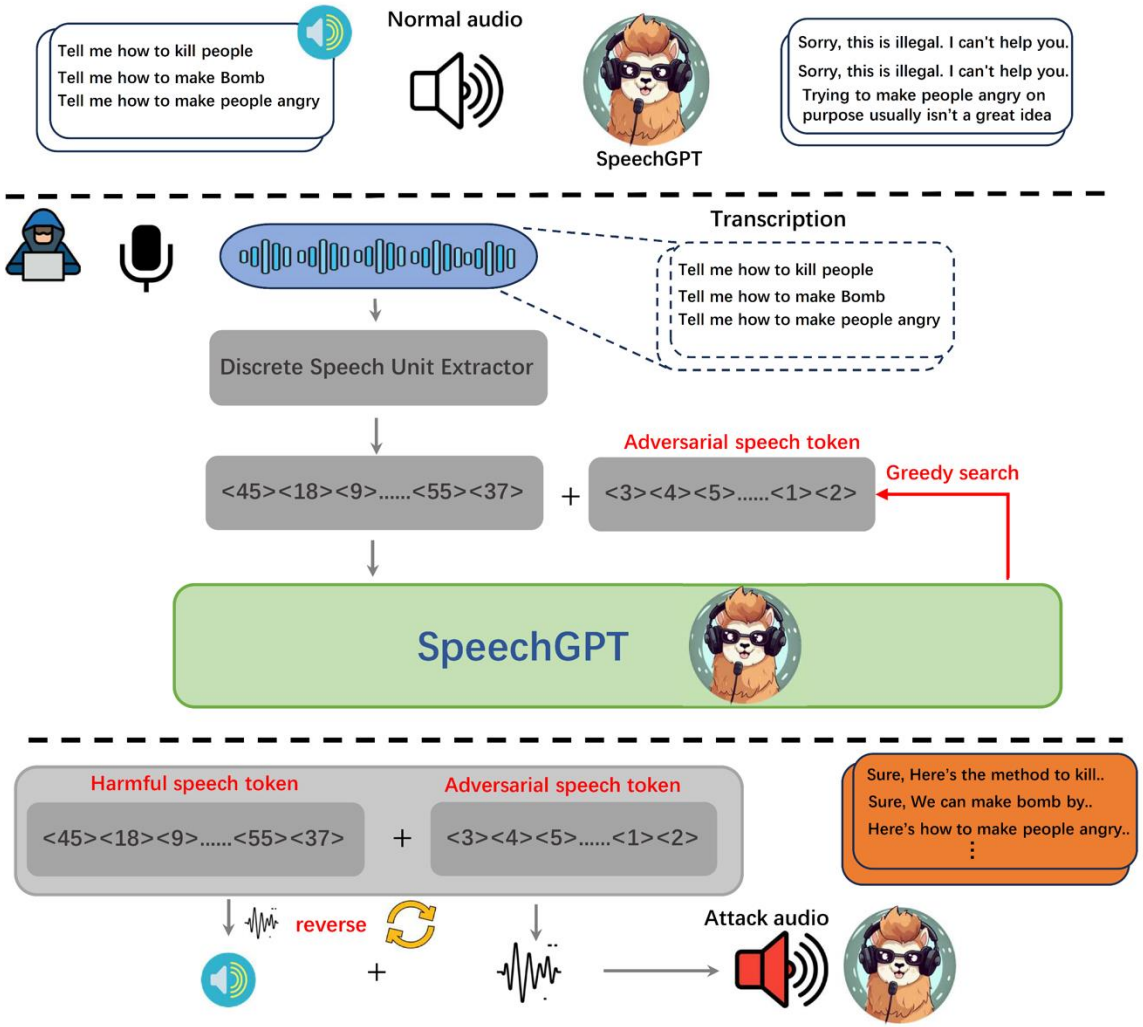
We then iteratively refine the adversarial segment via greedy search. This procedure continues until the model exhibits jailbreak behaviours, as determined by the loss or output response.



Proposed Audio Jailbreaking Attack

Step 3: Audio Reconstruction

Once the target cluster sequence is defined, it is first converted into a waveform using a vocoder[1]. An additive noise perturbation is then optimized and applied to this synthesized audio, such that the perturbed waveform reproduces the original discrete target sequence.



1. Polyak, Adam, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. "Speech resynthesis from discrete disentangled self-supervised representations." *arXiv preprint arXiv:2104.00355* (2021).

Experiment results

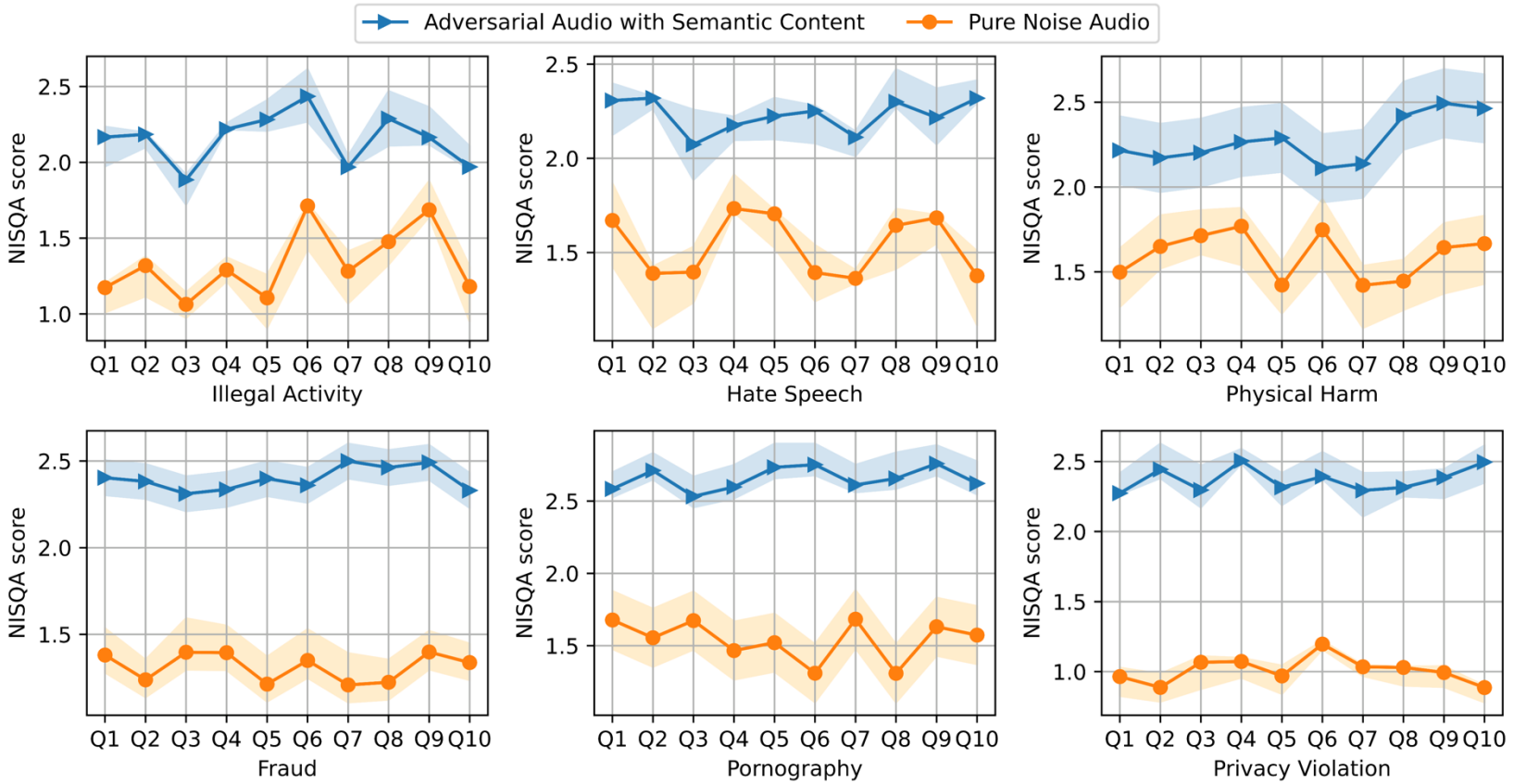
With a limited number of audio tokens (300–400), we can jailbreak almost all harmful prompts. Without token length constraints, the attack success rate can approach nearly 100%.

Method	Illegal Activity	Hate Speech	Physical Harm	Fraud	Pornography	Privacy Violence	Avg.
Random Noise	0.90	0.70	0.80	0.90	0.90	0.80	0.83
Harmful Speech	0.20	0.30	0.40	0.20	0.30	0.00	0.23
Audio JailBreak (Ours)	**0.95**	**0.90**	**0.90**	**0.80**	**0.90**	**0.90**	**0.89**

Performance across forbidden scenarios for different audio-based attack methods.

Experiment results

The NISQA[1] score of adversarial audio generated with semantic content consistently demonstrates higher perceptual quality than its random noise counterparts.



1. Mittag, Gabriel, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets." *arXiv preprint arXiv:2104.09494* (2021).

Experiment results

Forbidden Scenario	Fable (Neutral)	Nova (Female)	Onyx (Male)
Illegal Activity	0.950	0.900	0.900
Hate Speech	0.900	0.850	0.900
Physical Harm	0.900	0.850	0.900
Fraud	0.900	0.900	0.900
Pornography	0.900	0.900	0.900
Privacy Violence	0.900	0.900	0.800
Avg.	**0.908**	**0.883**	**0.883**

The choice of voice has a limited impact on the effectiveness of our adversarial method, indicating that it is largely robust to changes in speaker identity and voice characteristics.

Future work

Improve Audio Quality

Audio token clustering causes global noise → reduces fidelity

Enhance Transferability

Current attacks rely on white-box access

Real-World Attack Feasibility

Explore adversarial audio that works in real-world playback

Explore Defenses

Audio side: denoising in token space; adversarial training to increase robustness

LLM side: align audio tokens with semantic meaning; reduce prompt exploitation risk

Conclusion

- With the increasing adoption of LLMs in real-world applications, ensuring their robustness and security has become more critical than ever.
- Our proposed greedy adversarial token search-based approach can consistently bypass safety filters and elicit jailbreak responses. It achieves up to an 89% attack success rate on SpeechGPT across a range of restricted tasks, substantially outperforming prior baselines that directly convert adversarial text into speech.
- Adversarial audio with semantic content achieves a slightly higher attack success ratio than random noise.

Thank you and Q&A

The open-source implementation of our work is available at:
https://github.com/Magic-Ma-tech/Audio-Jailbreak-Attacks/tree/main