INTERSPEECH 2025 17-21 AUGUST - ROTTERDAM

TU Delft Delft University of Technology

Audio Demo LinkedIn

AdvSpeech: Adversarial Attack Against Zero-Shot Voice Cloning

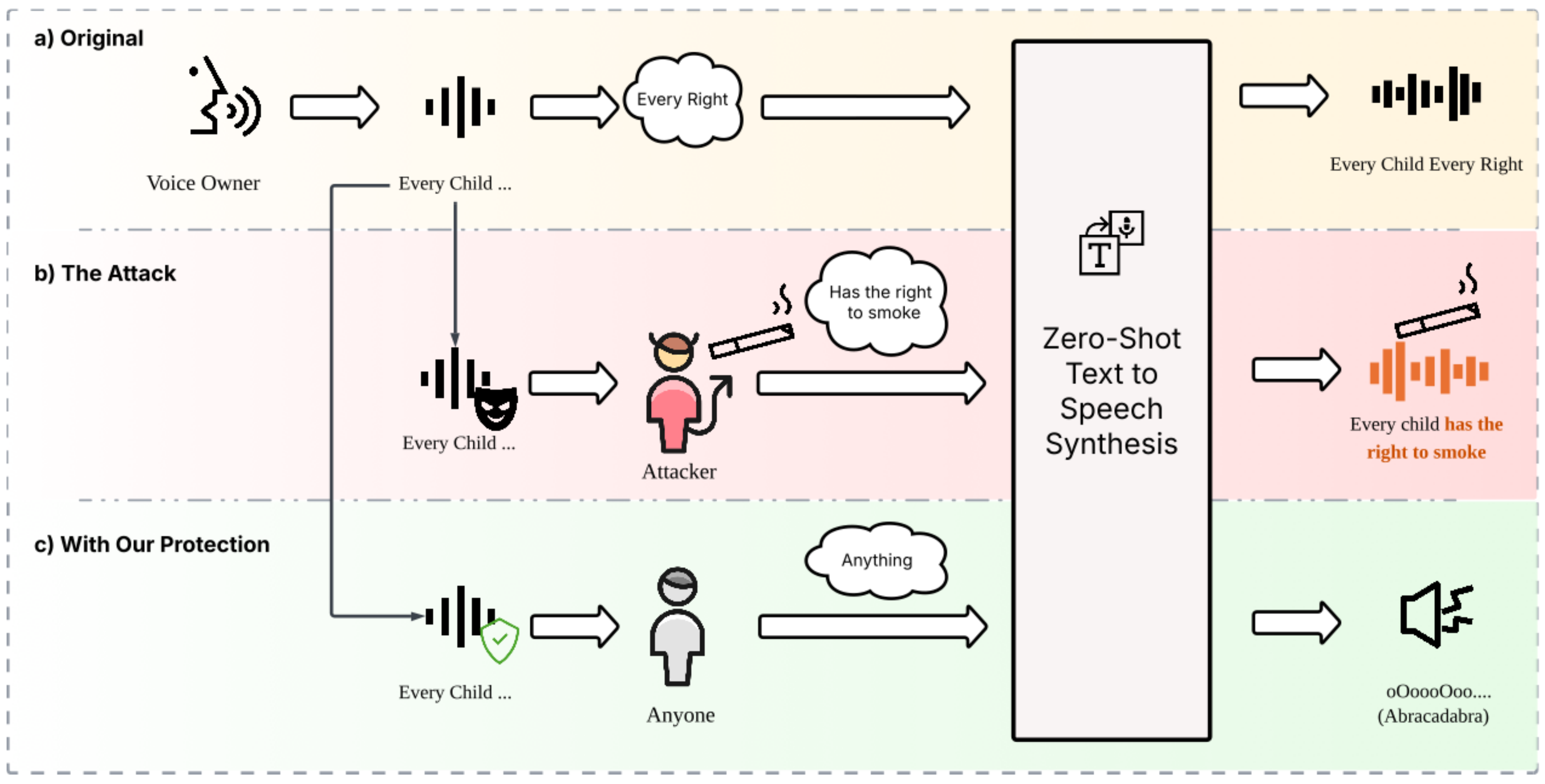
Renyi Yang{R.Yang-7@student.tudelft.nl}, Delft University of Technology

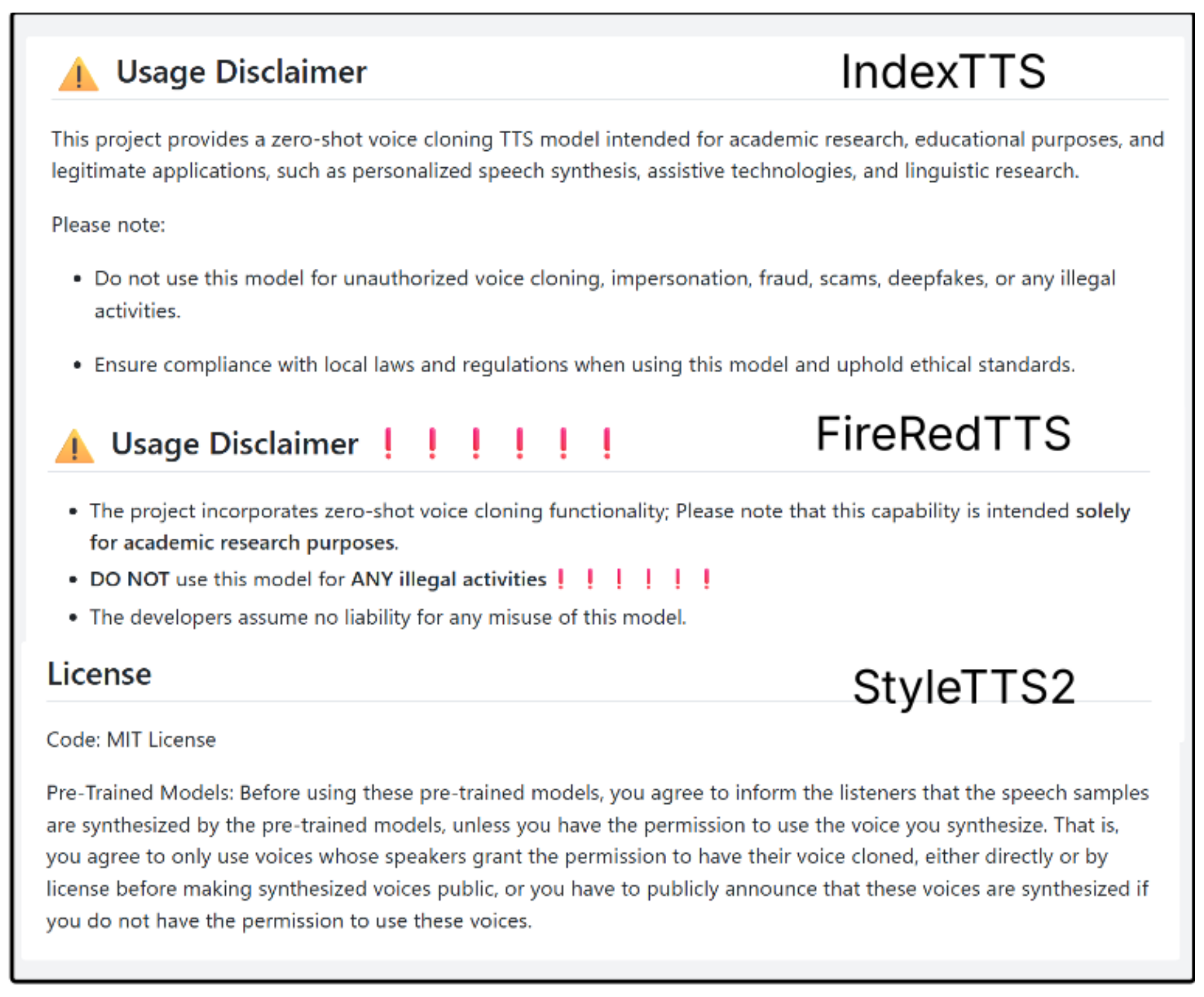
01 Background

Background Zero-shot TTS clones a voice from just seconds of audio at inference, which brings significant misuse risks.

Research Gap Watermarking and deepfake detection provide traceability but do not stop harmful content from being generated.

Proposed Solution A proactive, imperceptible and publisher-side protection that preserves perceived quality while nullifying zero-shot TTS outputs.





⚠ Usage Disclaimer IndexTTS

This project provides a zero-shot voice cloning TTS model intended for academic research, educational purposes, and legitimate applications, such as personalized speech synthesis, assistive technologies, and linguistic research.

Please note:

- Do not use this model for unauthorized voice cloning, impersonation, fraud, scams, deepfakes, or any illegal activities.
- Ensure compliance with local laws and regulations when using this model and uphold ethical standards.

⚠ Usage Disclaimer ❗ ❗ ❗ ❗ ❗ ❗ FireRedTTS

- The project incorporates zero-shot voice cloning functionality; Please note that this capability is intended **solely for academic research purposes.**
- **DO NOT** use this model for **ANY illegal activities** ❗ ❗ ❗ ❗ ❗ ❗
- The developers assume no liability for any misuse of this model.

License StyleTTS2

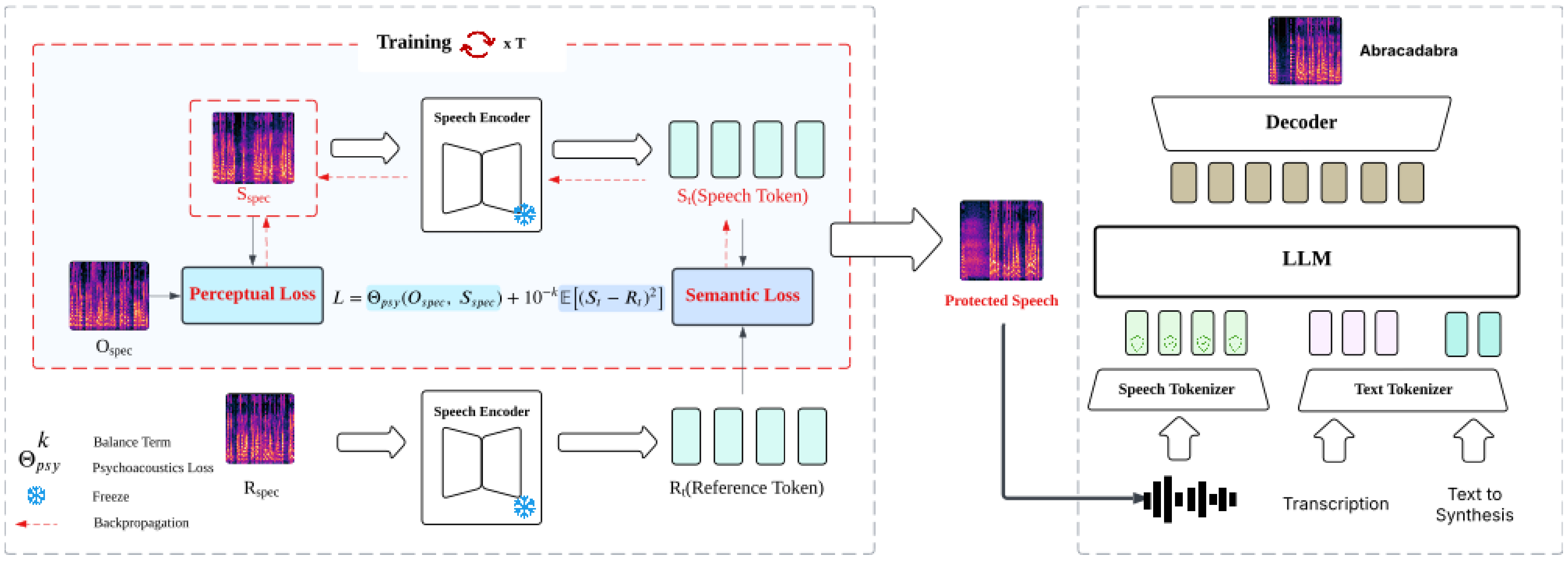
Code: MIT License

Pre-Trained Models: Before using these pre-trained models, you agree to inform the listeners that the speech samples are synthesized by the pre-trained models, unless you have the permission to use the voice you synthesize. That is, you agree to only use voices whose speakers grant the permission to have their voice cloned, either directly or by license before making synthesized voices public, or you have to publicly announce that these voices are synthesized if you do not have the permission to use these voices.

02 Methodology

On our training stage, we learn a tiny, imperceptible perturbation that guide the audio's speech-token sequence toward a reference speech while a psychoacoustic loss hides the change.

The protected audio can be shared as-is . On the attacker inference stage, when passed through zero-shot TTS, the token–transcript mismatch makes the system speak fluent but incorrect content.



$$L = \Theta_{psy}(O_{spec}, S_{spec}) + 10^{-k}\mathbb{E}[(S_t - R_t)^2]$$

1) Our Training Stage

2) Attaker Inference Stage

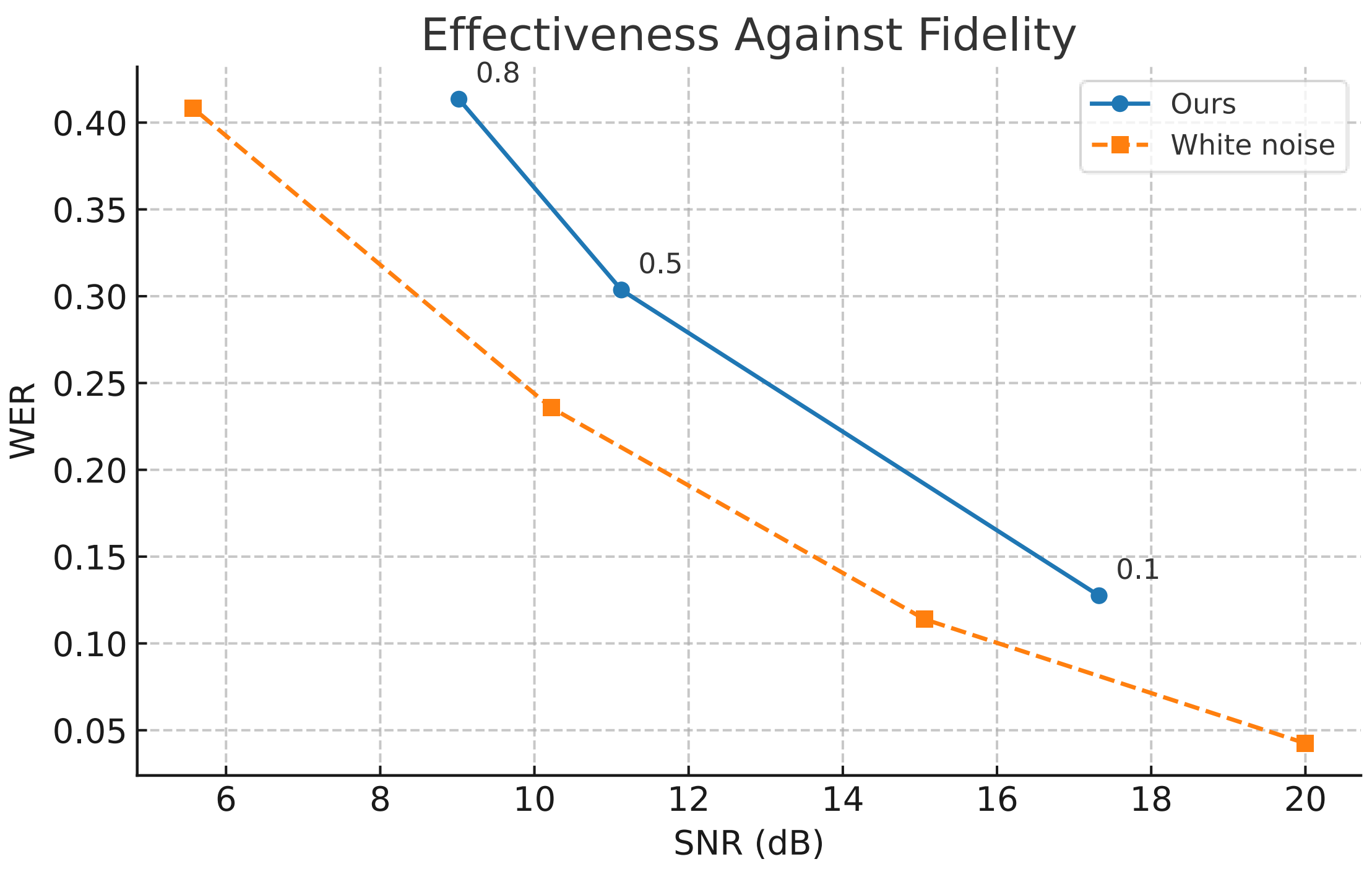
03 Experimental Results

Dataset	Method	Synth.	Effectiveness WER↑	Effectiveness WIL↑	BLEU↓	Fidelity SNR↑	Fidelity PESQ↑	Fidelity SECS↓
LibriTTS	Raw	Cosyvoice	0.03 ± 0.03	0.05 ± 0.05	0.96 ± 0.05	N	N	N
		Spark-TTS	0.01 ± 0.02	0.02 ± 0.04	0.97 ± 0.05			
		VALL-E	0.21 ± 0.14	0.33 ± 0.19	0.67 ± 0.18			
	Antifake[1]	Cosyvoice	0.09 ± 0.07	0.17 ± 0.12	0.84 ± 0.12	13.29 ± 3.54	1.27 ± 0.21	0.22 ± 0.06
		Spark-TTS	0.10 ± 0.14	0.17 ± 0.17	0.82 ± 0.17			
		VALL-E	0.73 ± 0.25	0.83 ± 0.16	0.20 ± 0.14			
	SafeSpeech[2]	Cosyvoice	0.10 ± 0.10	0.17 ± 0.14	0.82 ± 0.15	6.62 ± 3.59	1.08 ± 0.08	0.35 ± 0.06
		Spark-TTS	0.41 ± 0.34	0.52 ± 0.31	0.47 ± 0.29			
		VALL-E	1.12 ± 0.28	0.96 ± 0.07	0.07 ± 0.08			
	Ours	Cosyvoice	0.90 ± 0.37	0.85 ± 0.27	0.17 ± 0.27	18.33 ± 2.43	2.06 ± 0.22	0.08 ± 0.03
		Spark-TTS	0.28 ± 0.29	0.35 ± 0.29	0.66 ± 0.28	14.94 ± 1.70	1.34 ± 0.13	0.20 ± 0.06
		VALL-E	0.87 ± 0.26	0.93 ± 0.09	0.10 ± 0.09	11.41 ± 2.30	1.19 ± 0.12	0.21 ± 0.06

Performance Comparison Between the Proposed Method and the Baseline Methods

We evaluate the attack effectiveness by comparing the cloned speech and the GT transcript. Fidelity is measured on the adversarial examples related to the raw samples.

Cosyvoice is cracked, minimal distortion can lead to completely meaningless output(~0.9 WER). Other synthesizers show different level of resistance on our method.



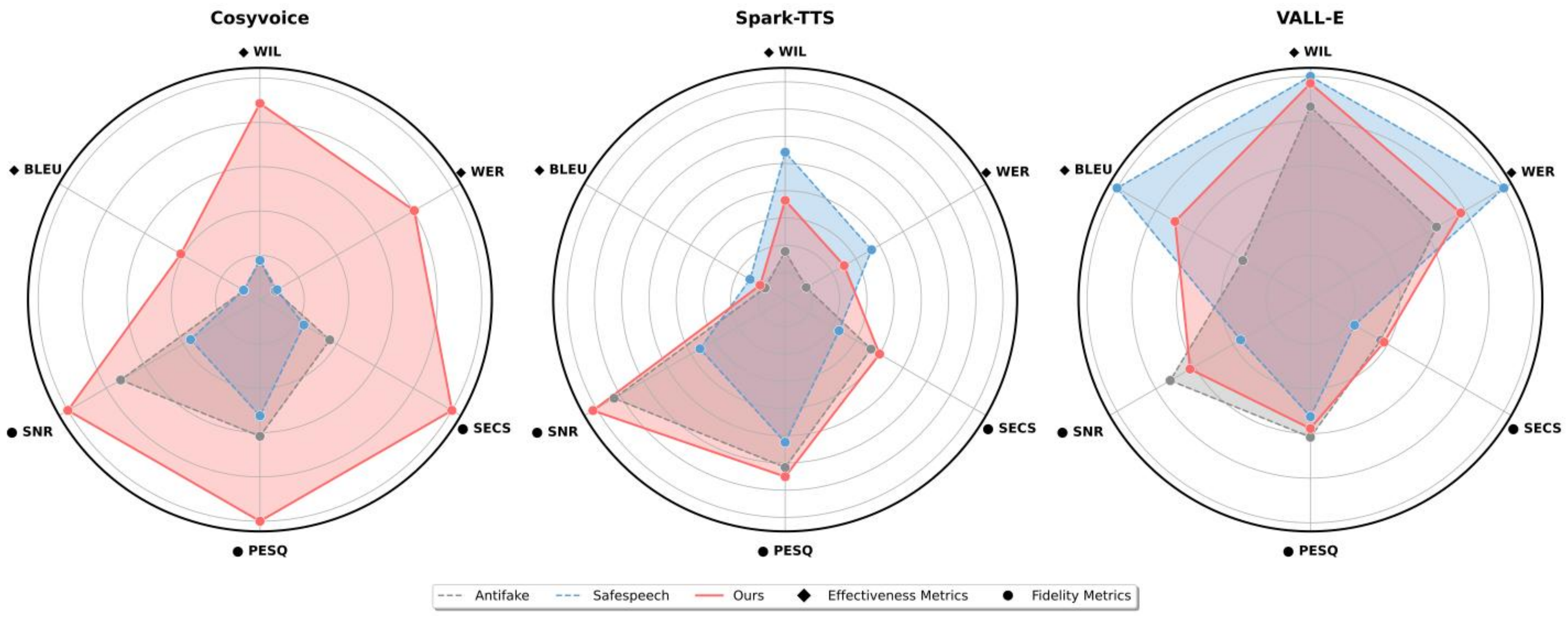
Ours: same SNR → ↑WER

Our attack crafts small and model-aware perturbations rather than injecting untargeted white noise.

This reveals that LM-based TTS systems are not robust to targeted adversarial perturbations.

04 Analyses and Conclusions

Different TTS system shows a significant difference in the level of noise resistance and the robustness the adversarial method.



We propose ***AdvSpeech*** , an adversarial attack method that aggressively prevents zero-shot TTS from producing an intelligible speech corresponding to any given text prompt.

Our adversarial examples introduce minimal audible distortion compared to existing methods, while simultaneously delivering a higher overall success rate.

Reference

[1]. Yu, Zhiyuan, Shixuan Zhai, and Ning Zhang. "Antifake: Using adversarial audio to prevent unauthorized speech synthesis." Proceedings of the 2023 *ACM SIGSAC Conference on Computer and Communications Security*. 2023.
[2]. Tan, Xingwei, Lyu, Chen, Umer, Hafiz Muhammad, Khan, *et.al.*"SafeSpeech: A Comprehensive and Interactive Tool for Analysing Sexist and Abusive Language in Conversations." *arXiv*:2503.06534. 2025.