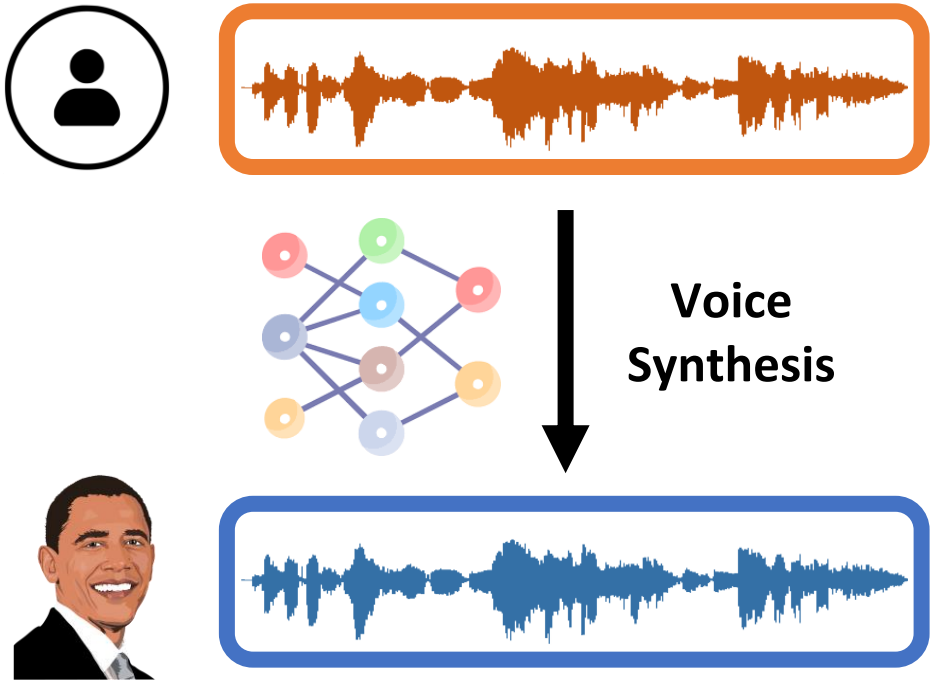
VSMask: Defending Against Voice Synthesis Attack via Real-Time Predictive Perturbation

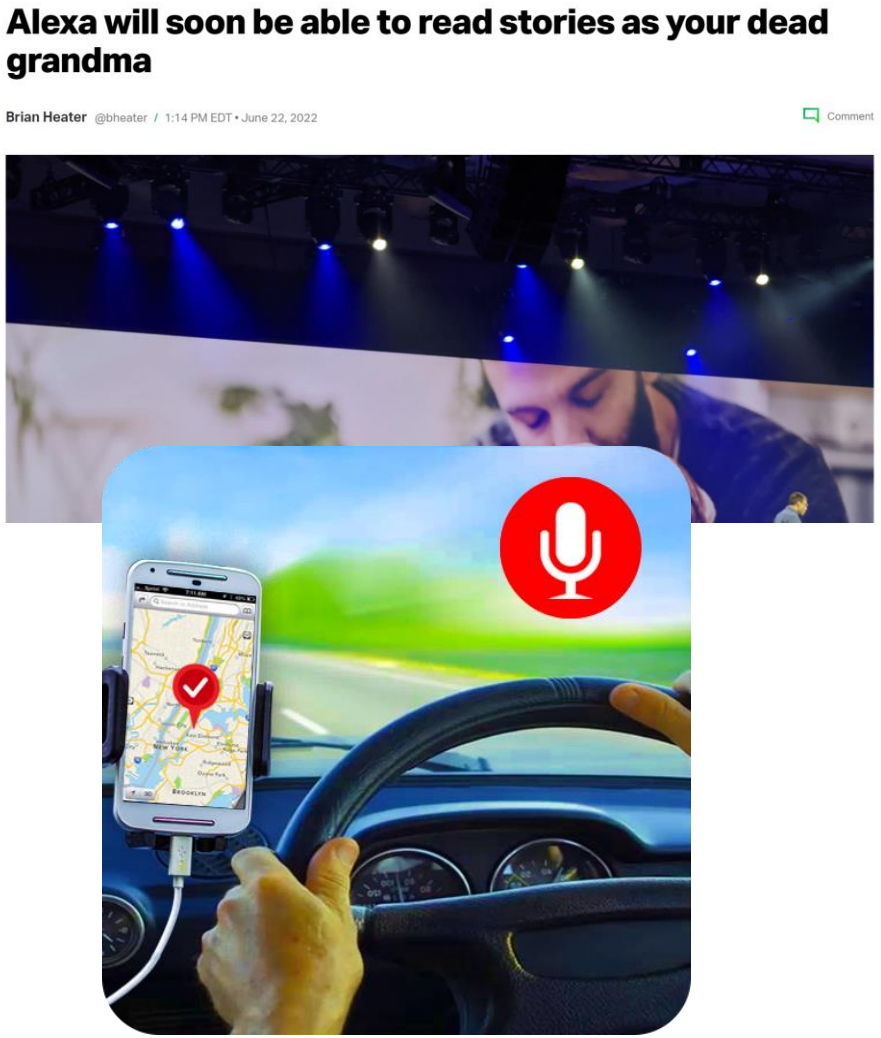
Yuanda Wang, Hanqing Guo, Guangjing Wang, Bocheng Chen, Qiben Yan

SEIT Lab

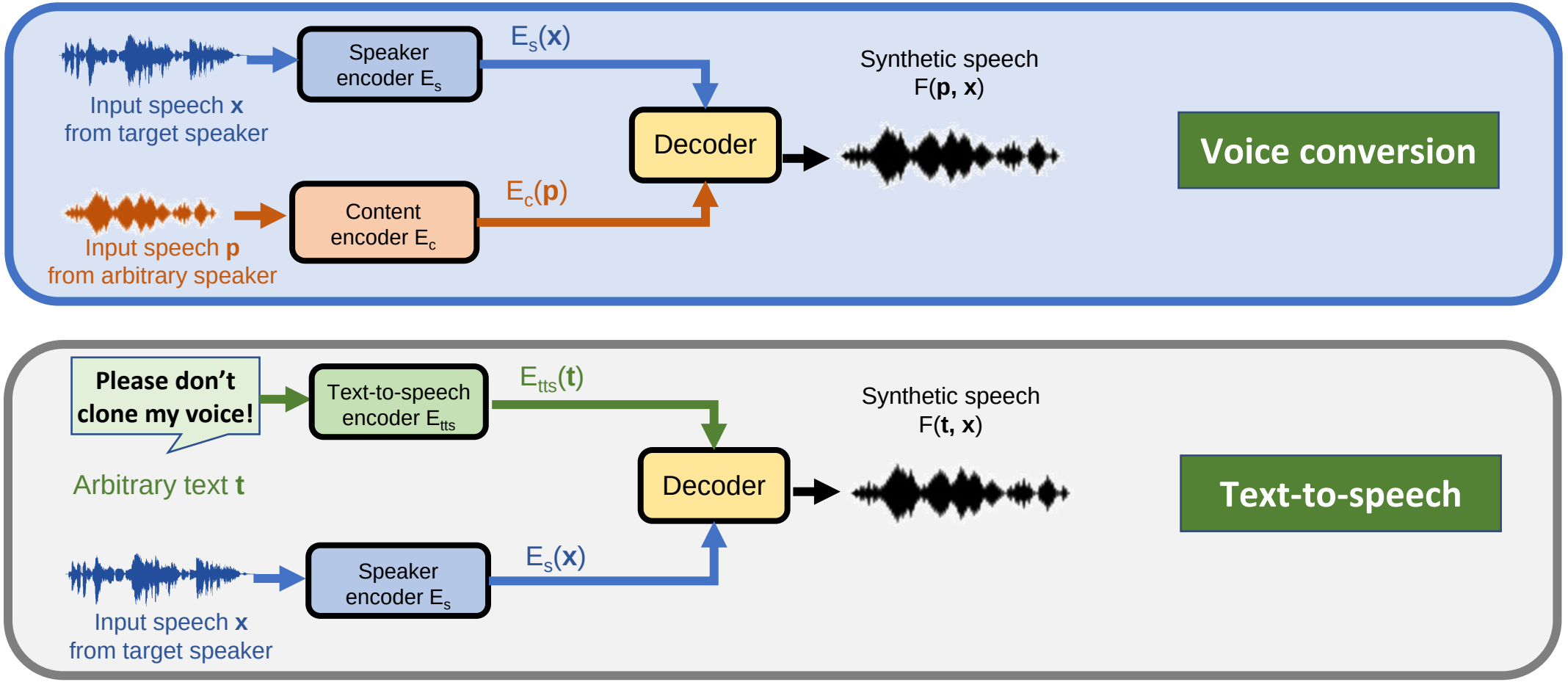
Michigan State University

Voice Synthesis





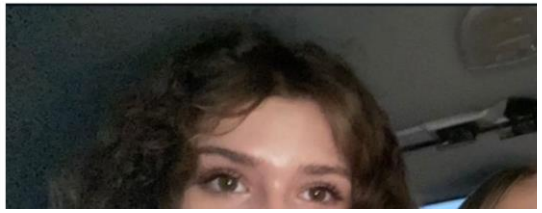
Voice Synthesis Methods



Deepfake Voice Threats

‘Mom, these bad men have me’: She believes scammers cloned her daughter’s voice in a fake kidnapping

By Faith Karimi, CNN
Updated 9:26 AM EDT, Sat April 29, 2023



ARTIFICIAL INTELLIGENCE · Published April 18, 2023 9:00am EDT

Scammers use AI to clone voice, terrify family with fa call: 'Worst day of my

An Arizona mom said the AI voice-cloning of her daughter was 'awful and v

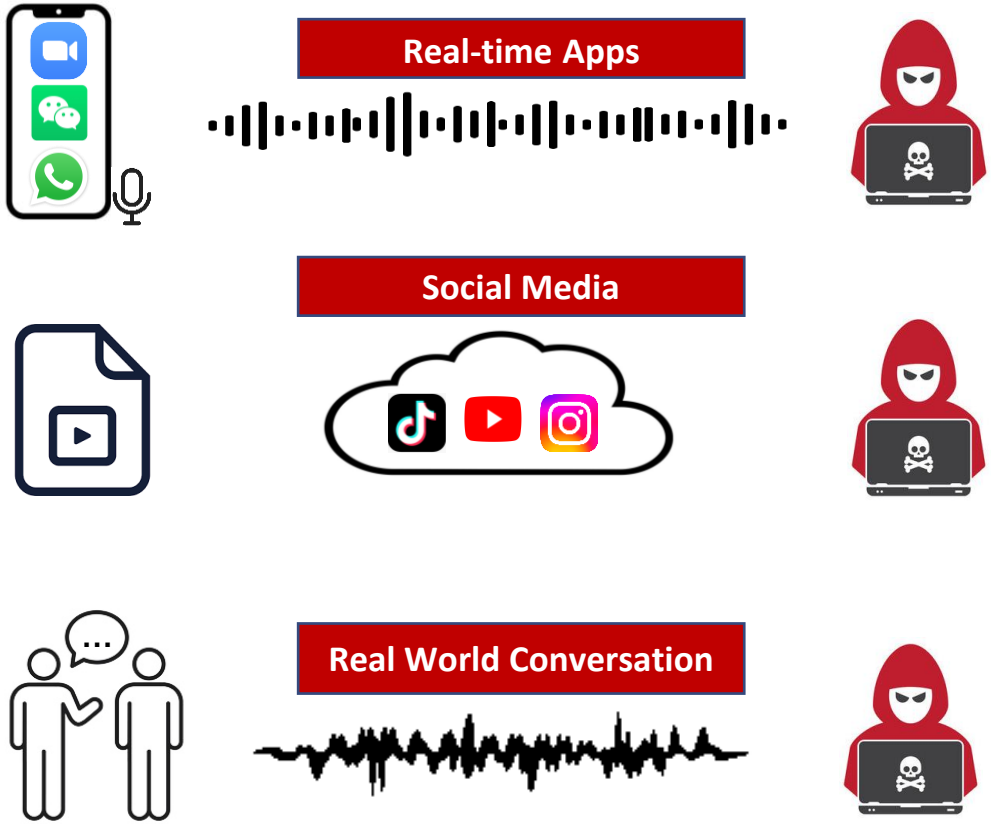
By Bailee Hill | Fox News

Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case

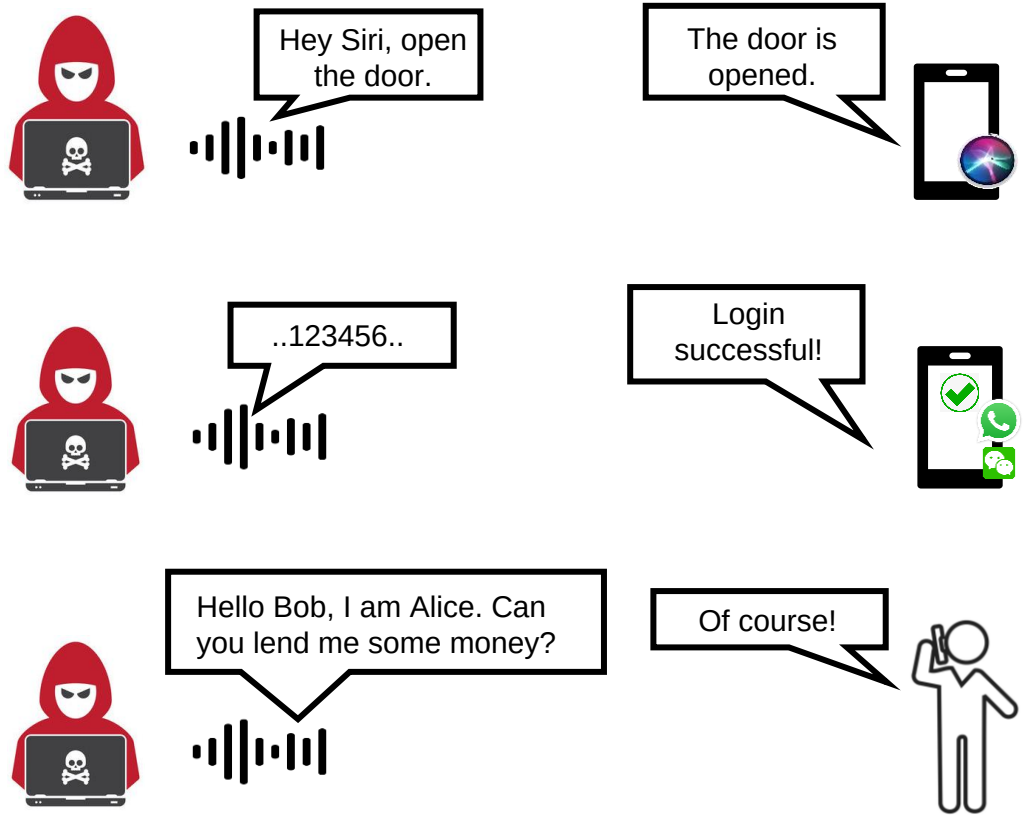
Scams using artificial intelligence are a new challenge for companies

https://www.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html
https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
https://www.foxnews.com/media/scammers-ai-clone-womans-voice-terrify-family-fake-ransom-call-worst-day-life

Threat Model

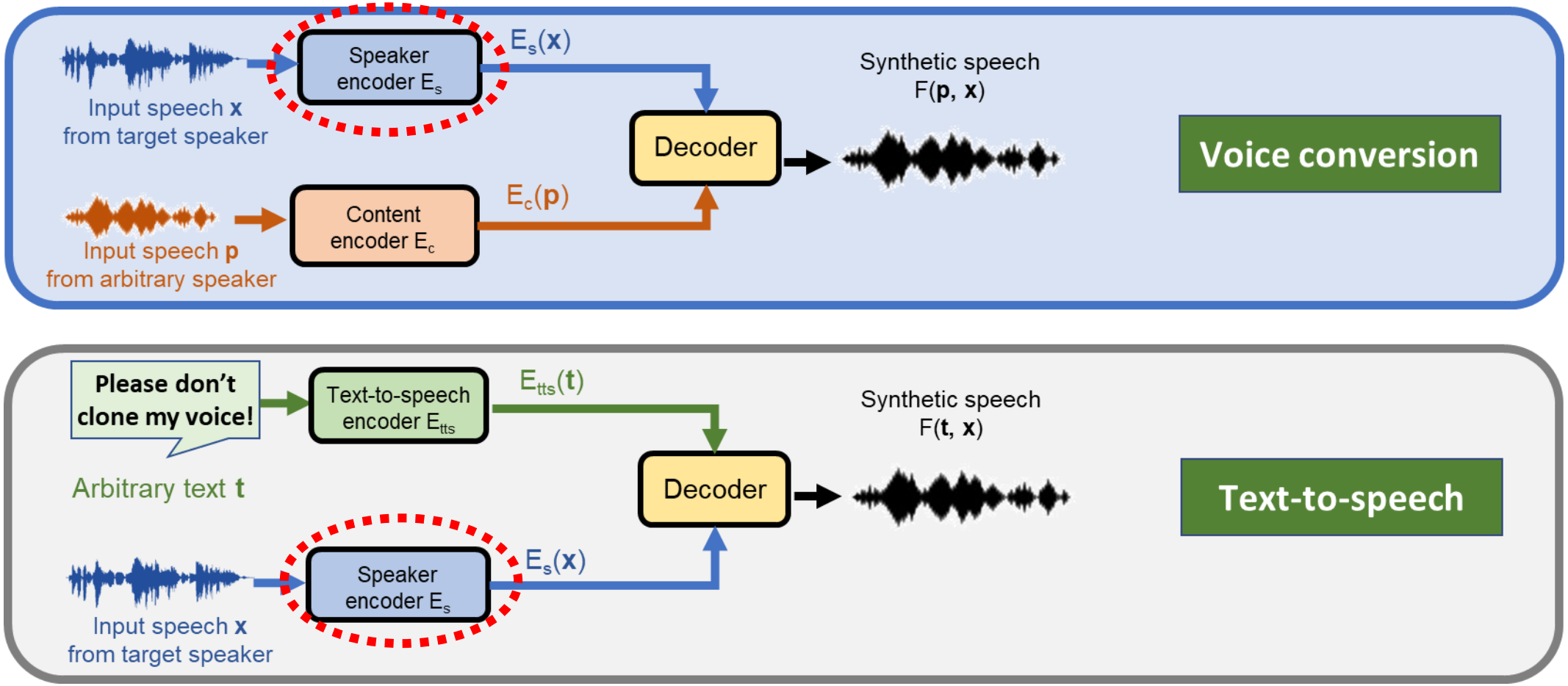


The adversary can hack your voice everywhere!

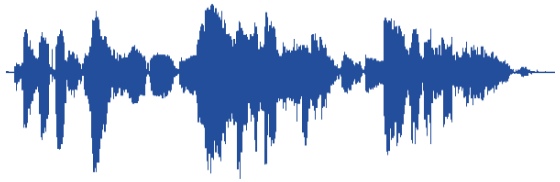


The synthetic voice can spoof both AI and human ears!

Defense against Voice Synthesis

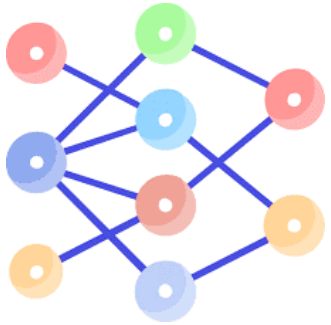


Defense against Voice Synthesis -Cont.

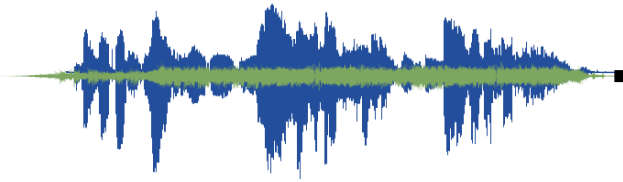


Clear speech sample $\mathbf{x}$

Speaker encoder E_s



$F(\mathbf{x})$ sounds like x



Adversarial sample $\mathbf{x} + \delta$

Speaker encoder E_s

$F(\mathbf{x} + \delta)$ sounds like y

$$\underset{\delta}{\text{minimize}} \quad \mathcal{L}\big(E_s(\mathbf{x}+\delta), E_s(\mathbf{y})\big) - \lambda\mathcal{L}\big(E_s(\mathbf{x}+\delta), E_s(\mathbf{x})\big)$$
$$\text{subject to} \quad \|\delta\|_\infty < \varepsilon$$

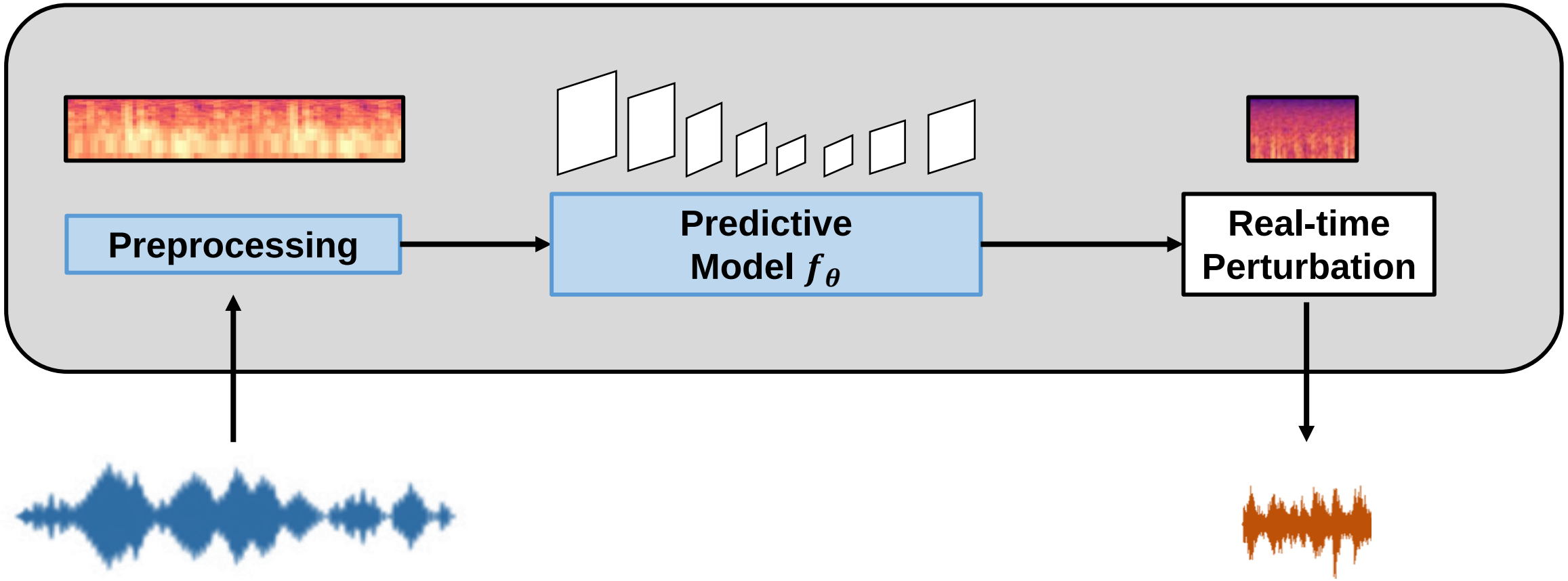
Challenges

- Existing defense cannot provide real-time protection.
- It is time-consuming to generate protected speech by gradient descent.
- There is perceptible noise in the protected audio.

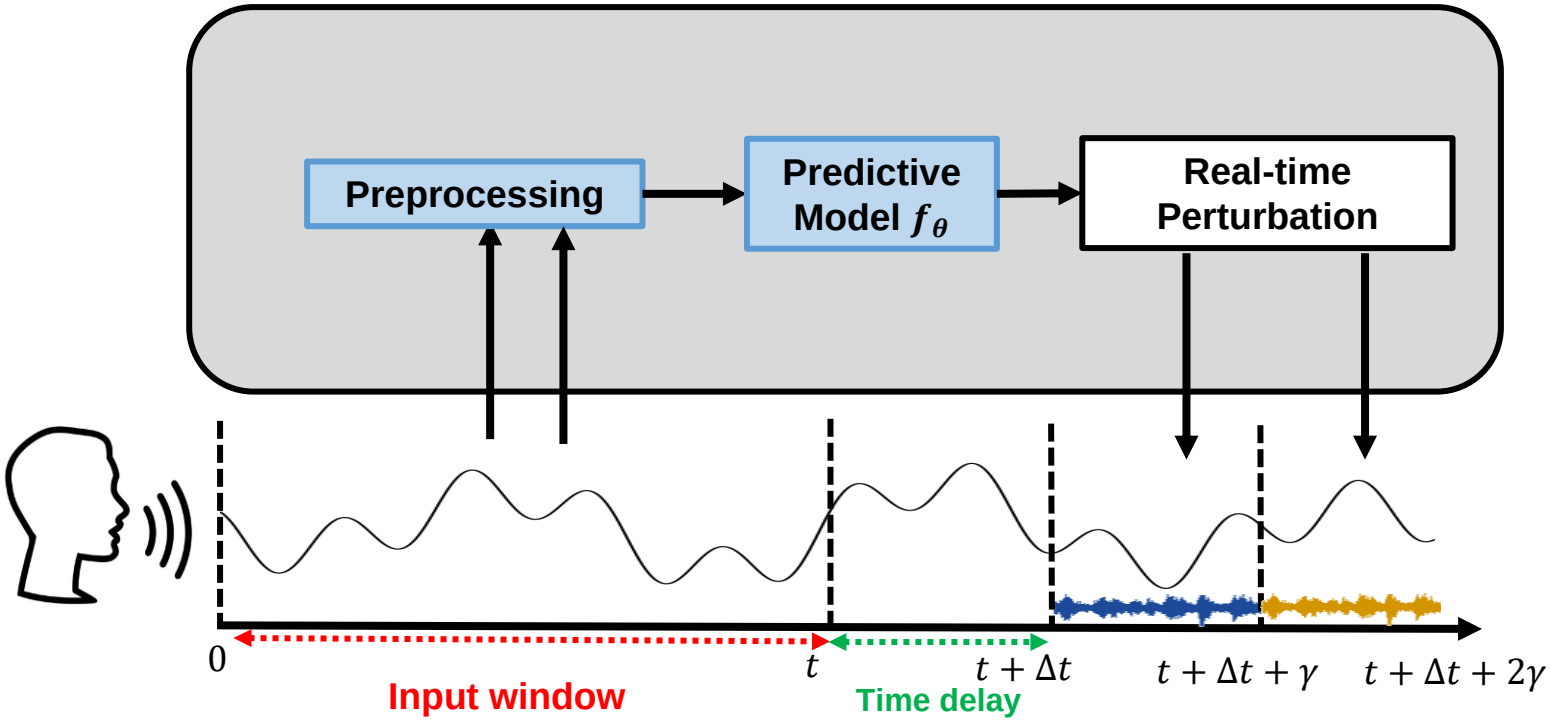
How to protect our voice in real-time without compromising audio quality ?

Predictive Model

- We can forecast the perturbation for upcoming live speech.



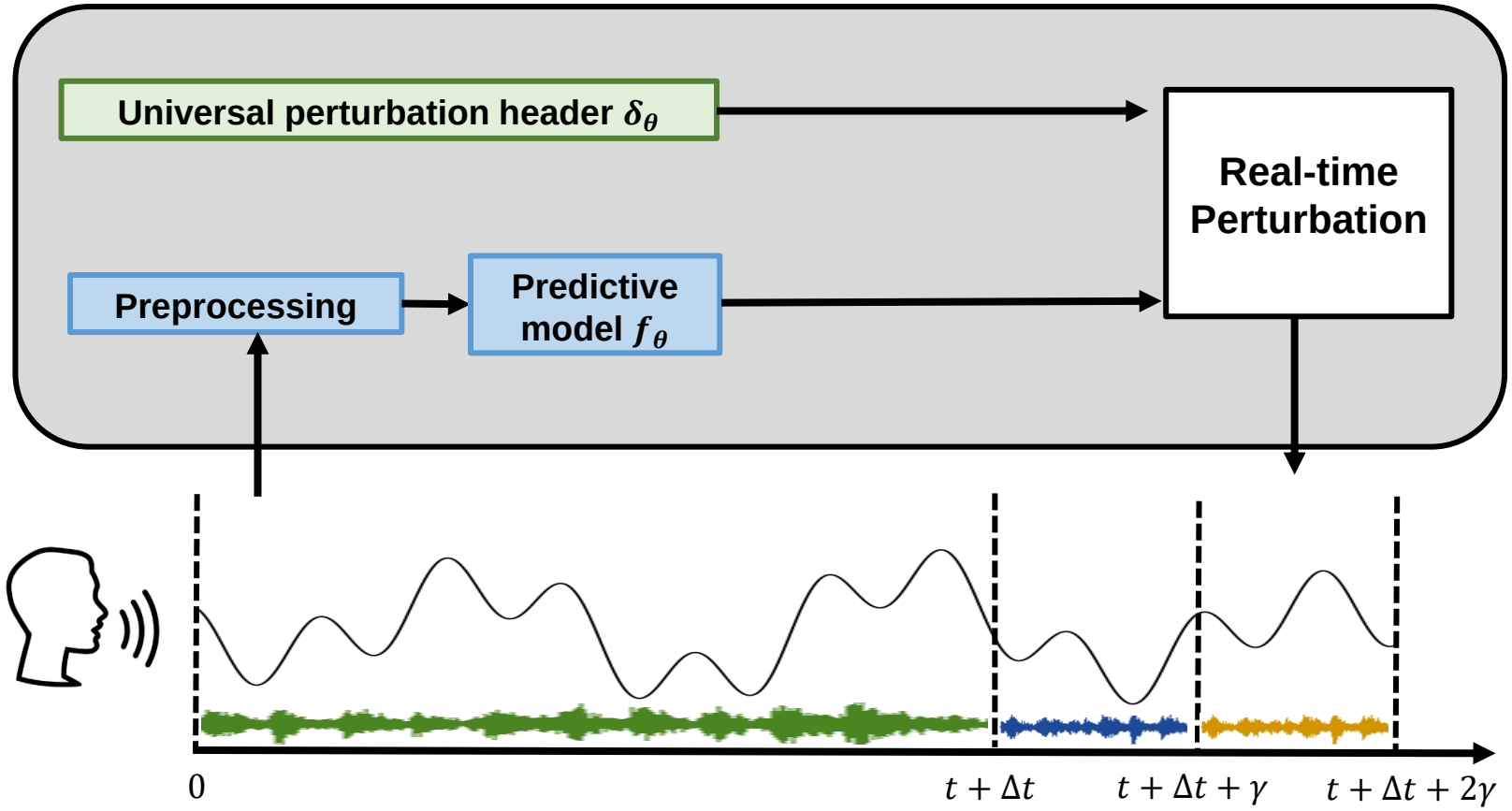
Predictive Model -Cont.



$$\underset{\theta}{\text{minimize}} \quad \mathcal{L}\big(\mathrm{E}_s(\mathbf{x}+\delta), \mathrm{E}_s(\mathbf{y})\big) - \lambda\mathcal{L}\big(\mathrm{E}_s(\mathbf{x}+\delta), \mathrm{E}_s(\mathbf{x})\big)$$

$$\text{subject to} \quad \delta_{t+\Delta t+\gamma} = f_\theta(\mathbf{x}_t) \quad \text{and} \quad \| \delta \|_\infty < \varepsilon$$

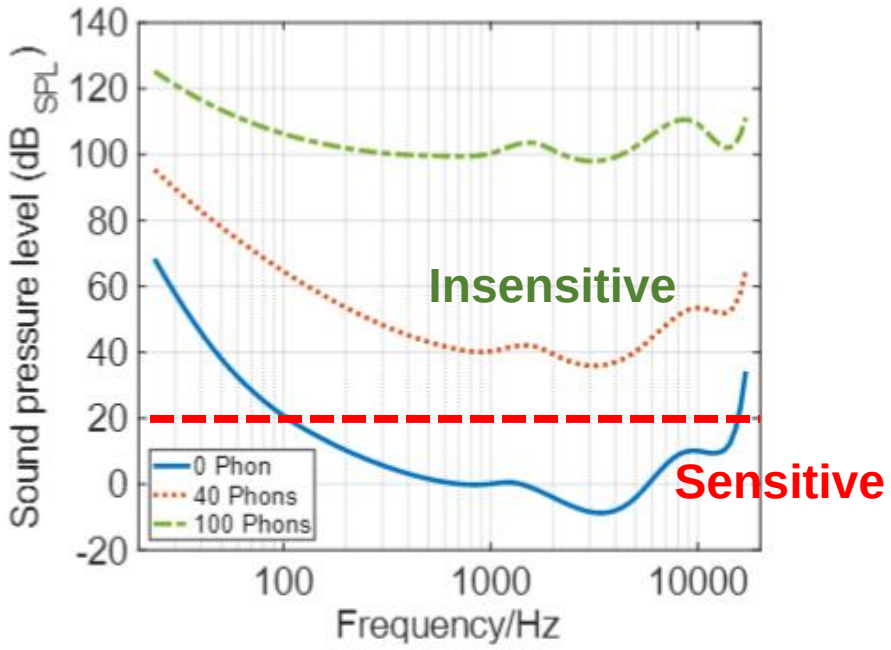
Universal Perturbation Header



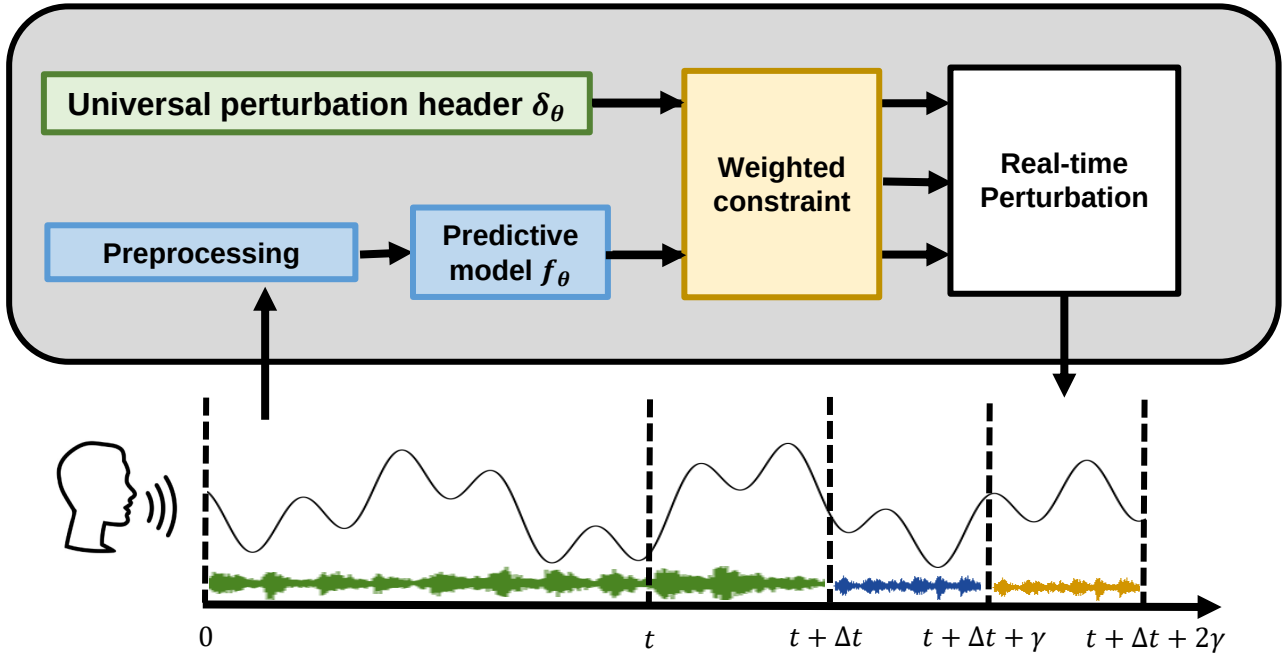
$$\underset{\delta_h}{\text{minimize}} \sum_{\mathbf{x}_i \in \mathcal{D}} \mathcal{L}_s(\delta_h, \mathbf{x}_i, \mathbf{y}), \text{ subject to } \| \delta \|_\infty < \varepsilon,$$

$$\text{and } \mathcal{L}_s(\delta_h, \mathbf{x}_i, \mathbf{y}) = \mathcal{L}\big(\mathrm{E}_s(\mathbf{x}_i + \delta), \mathrm{E}_s(\mathbf{y})\big) - \lambda \mathcal{L}\big(\mathrm{E}_s(\mathbf{x}_i + \delta), \mathrm{E}_s(\mathbf{x}_i)\big)$$

Perceptibility Mitigation

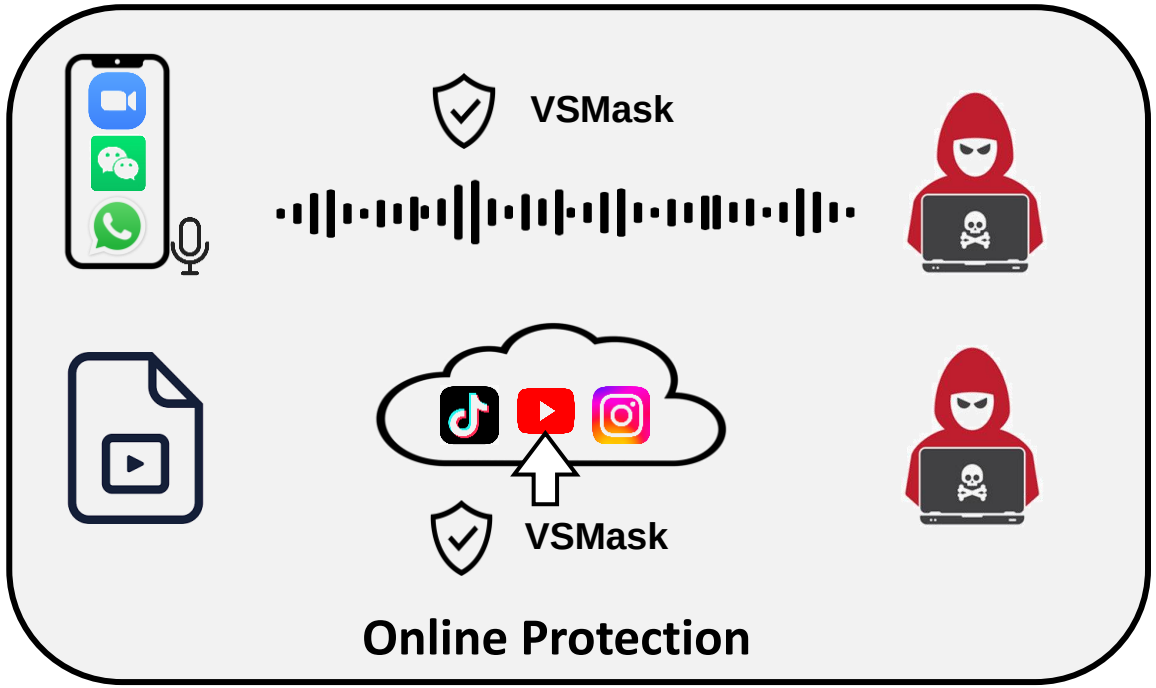


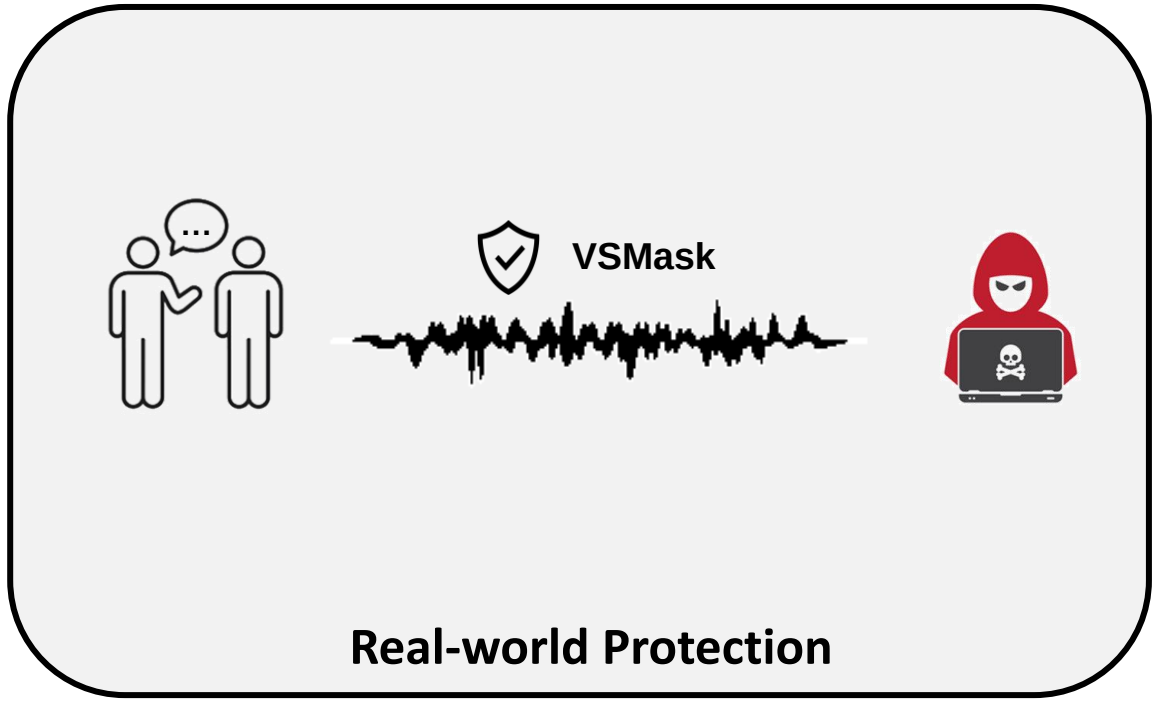
Human ears are more sensitive to 1.6 kHz to 4 kHz audio!



$$\delta = [\delta_{low} \quad \delta_{mid} \quad \delta_{high}]^T,$$
$$\text{subject to} \quad \| \delta_{low} \|_\infty < \varepsilon_1, \quad \| \delta_{mid} \|_\infty < \varepsilon_2, \quad \| \delta_{high} \|_\infty < \varepsilon_3$$

VSMask Application Scenarios





Real-time protection ✓

Zero latency ✓

Low perceptibility ✓

Evaluation Setup

Target Models

- AdaIN-VC (2019)
- AutoVC (2020)
- SV2TTS (2018)

Datasets

- VCTK Corpus (Voice Conversion)
- LibriSpeech (Text-to-speech)

Baseline Methods

- Random Noise
- Periodical Perturbation
- Online PGD
- Offline PGD

Parameters

- $t = 1.25$s (Input length)
- $\Delta t = \gamma = 0.4$s (Delay and output)
- $\varepsilon_{low} = 0.115, \varepsilon_{high} = 0.10$, $\varepsilon_{mid} = 0.085$

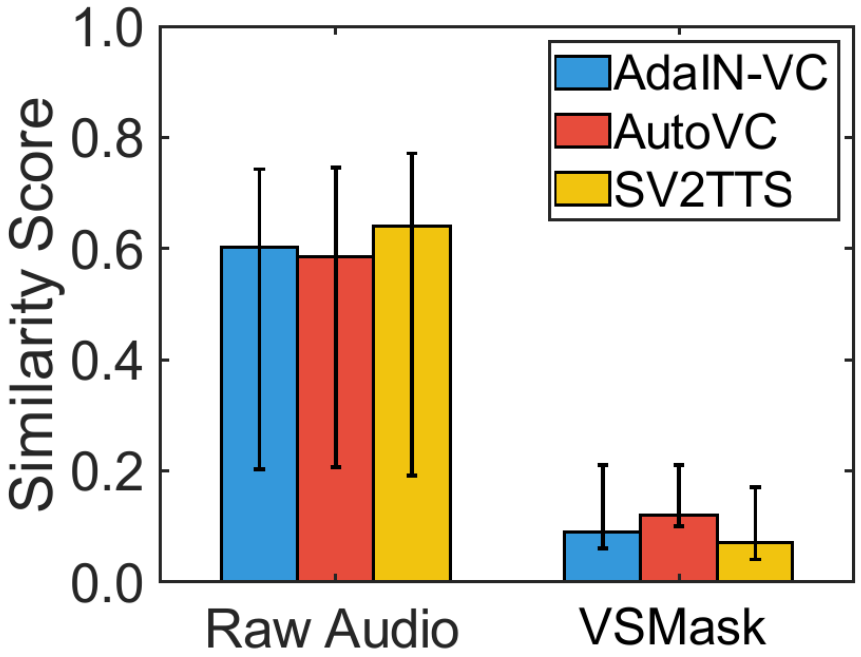
Evaluation on ASV

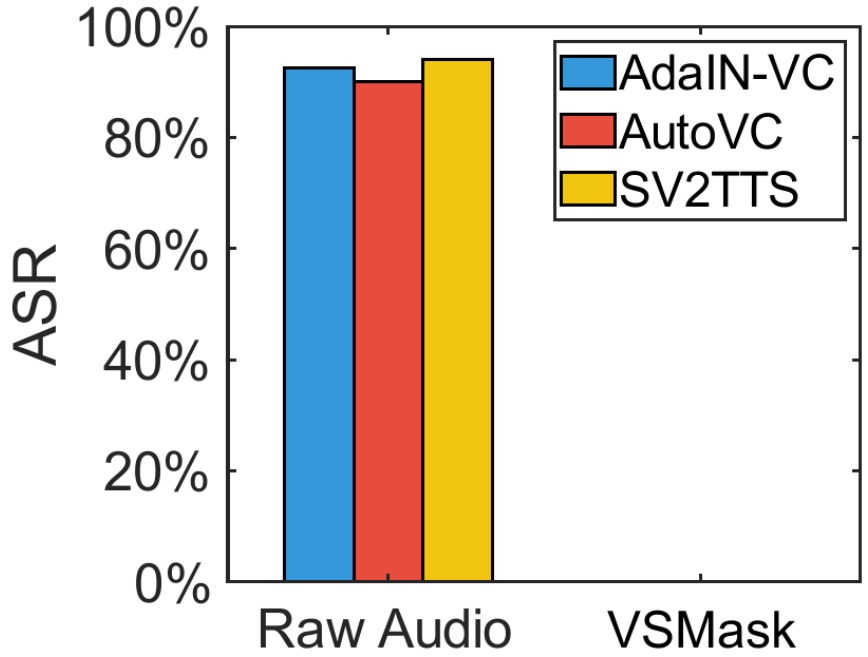
- We use SpeechBrain for speaker verification. (Threshold = 0.25)

Method	Male-to-Male		Female-to-Female		Male-to-Female		Female-to-Male	
	Score	ASR	Score	ASR	Score	ASR	Score	ASR
Raw speech	0.595	91.9%	0.612	93.2%	0.561	88.3%	0.546	86.0%
Random noise	0.516	86.6%	0.538	89.0%	0.505	84.0%	0.473	81.5%
Periodical Perturbation	0.192	11.0%	0.203	12.5%	0.177	9.8%	0.156	8.6%
Offline PGD	0.064	0.0%	0.085	0.0%	0.049	0.0%	0.055	0.0%
VSMask	**0.077**	**0.0%**	**0.104**	**0.0%**	**0.056**	**0.0%**	**0.073**	**0.0%**

We apply VSMask to defend against AdaIN-VC voice synthesis model. It outperforms all real-time defenses and achieves similar performance as offline PGD method.

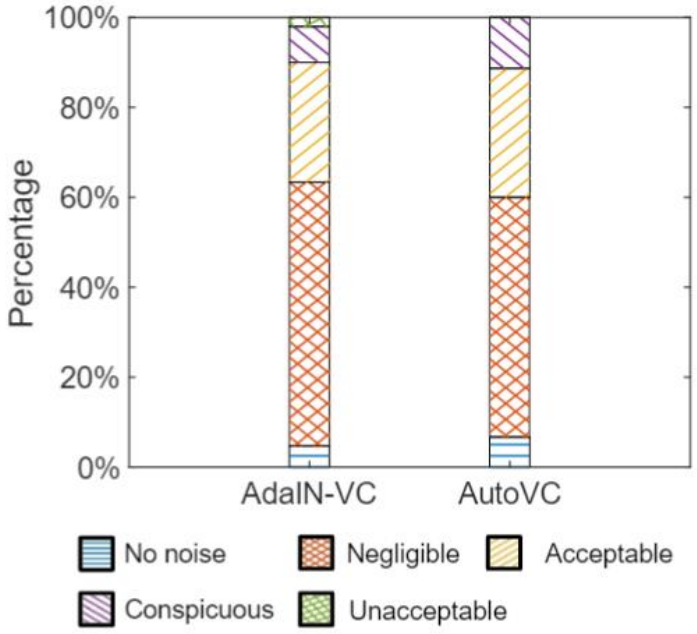
Evaluation on Different Models

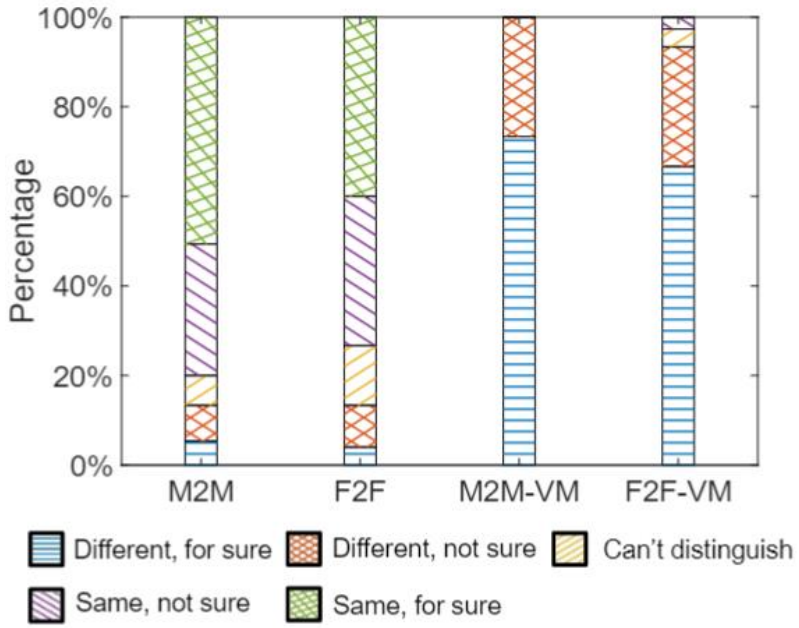


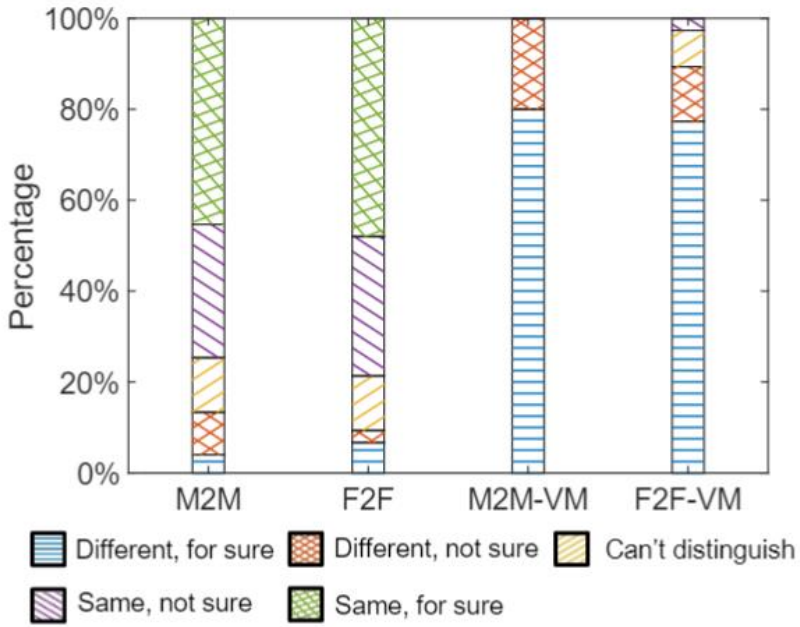


VSMask successfully defenses 3 different voice synthesis models. None of the synthetic speech samples can bypass the speaker verification.

Human Study







The perturbation is almost imperceptible for human ears!

The synthetic speech from protected samples can NEVER fool human ears!

Cross-model & Adaptive Attack Evaluation

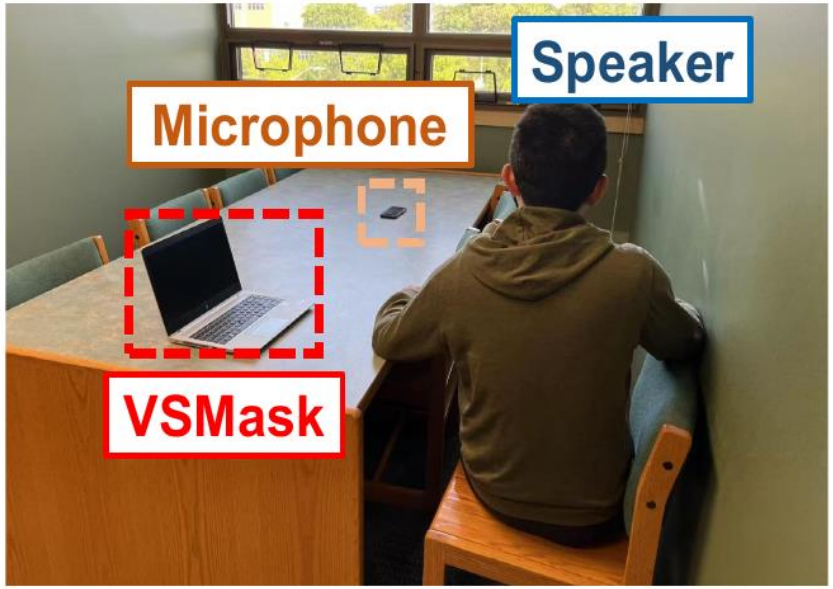
Source / Target	AdaIN-VC	AutoVC	SV2TTS
AdaIN-VC	--	15.0%	10.5%
AutoVC	12.8%	--	0.0%
SV2TTS	7.3%	15.2	--

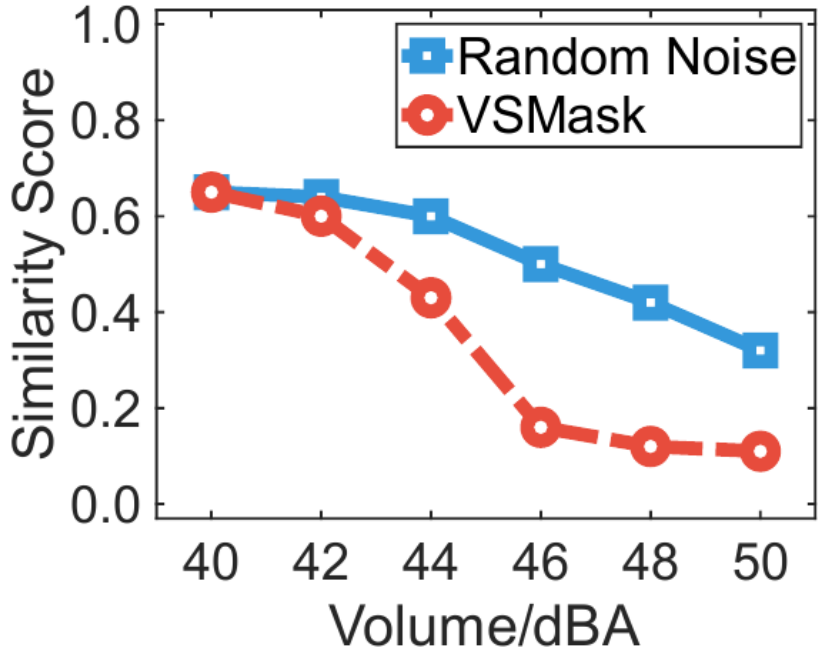
Adaptive methods	**None**	**Denoiser**	**WaveGuard**			
			Down-up (f=24k)	Quan-Dequan	Mel. (Bin=128)	LPC (Ord.=10)
Score	0.096	0.090	0.078	0.082	0.080	0.073

- Different input dimensions.
- Different training data.
- Different sampling rates.

- The perturbation is mel-spectrogram.
- Low audio quality degrades the performance.

Real-world Demonstration





VSMask can also protect our voice in physical world scenarios.

Discussion and Limitations

- **Discussion**

❑ **Real-time feasibility**

❑ **Adversarial training**

- **Limitations**

❑ **Powerful attackers**

❑ **Black-box defense**

❑ **Physical-world protection**

Conclusion

- We propose VSMask, a real-time defense mechanism against voice synthesis attacks based on predictive model.
- We optimize a universal perturbation header to indiscriminately protect speech with different lengths and sizes.
- We evaluate VSMask on three different voice synthesis models. The experimental results show that VSMask can provide real-time defense on both digital and physical spaces.

Scan me to check our website!

SEIT

https://seit.egr.msu.edu/

MICHIGAN STATE UNIVERSITY