Adversarial Music: 𝄞
Real world audio adversary against wake-word detection systems

Juncheng B. Li[♤], **Shuhui Qu**[◇], **Xinjian Li**[♤],

Joseph Szurley[♧], **Zico Kolter**[♤,♧], **Florian Metze**[♤]

♤ *Carnegie Mellon University*

♧ *Bosch Center for Artificial Intelligence*

◇ *Stanford University*

Motivation

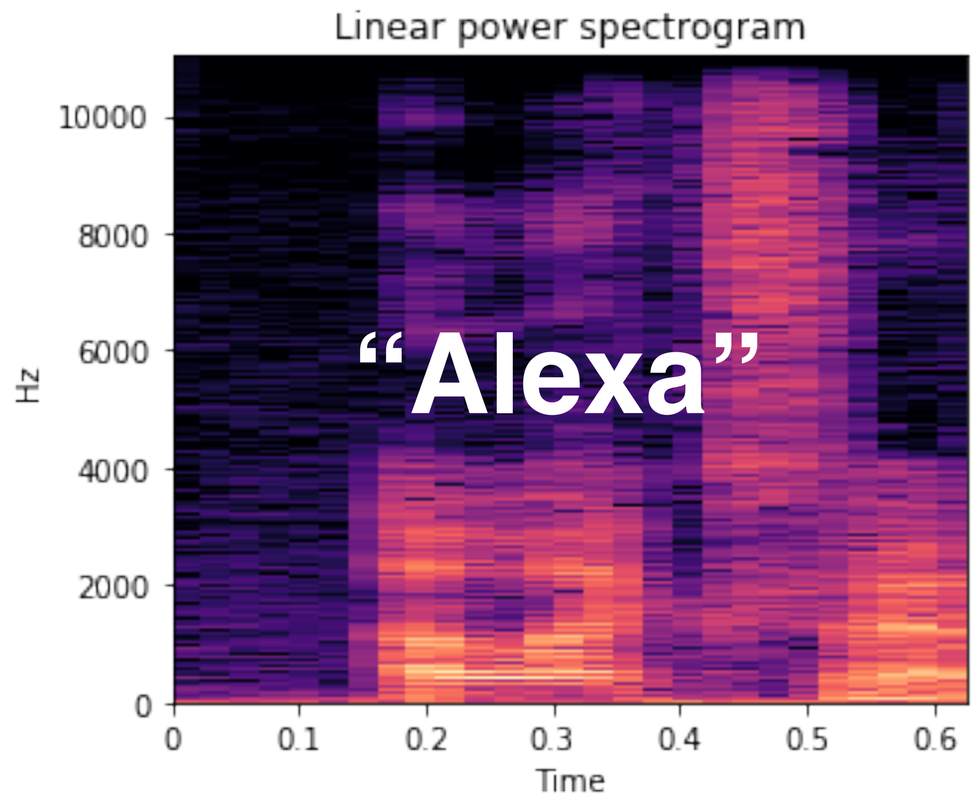
Adversarial Attack

not just a problem in vision

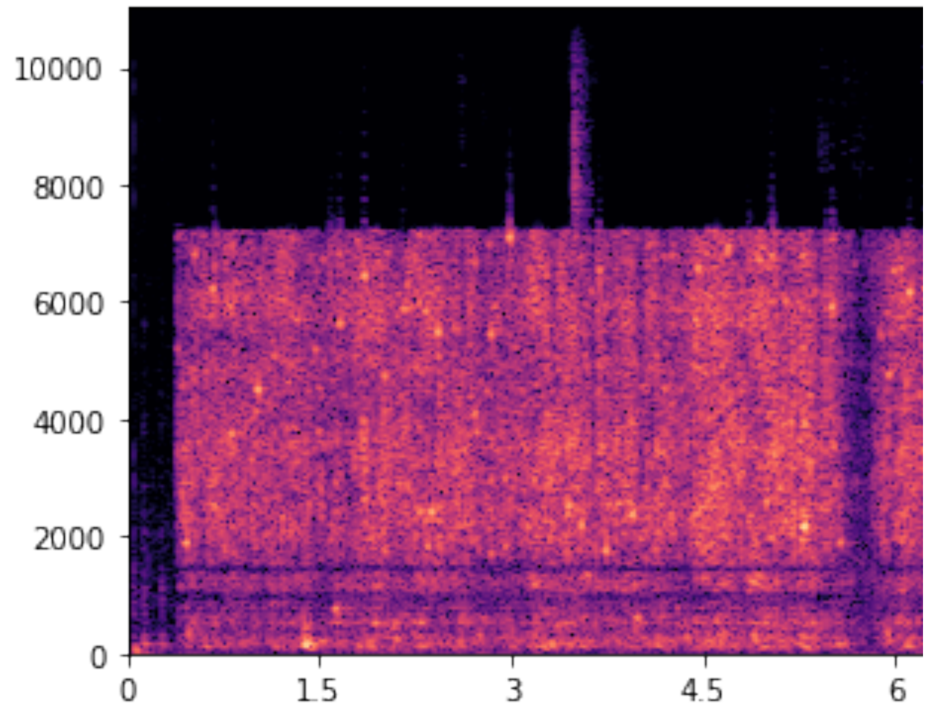
Li et al. [2019]

Existing audio attacks against Automatic Speech Recognition systems

not robust over-the-air

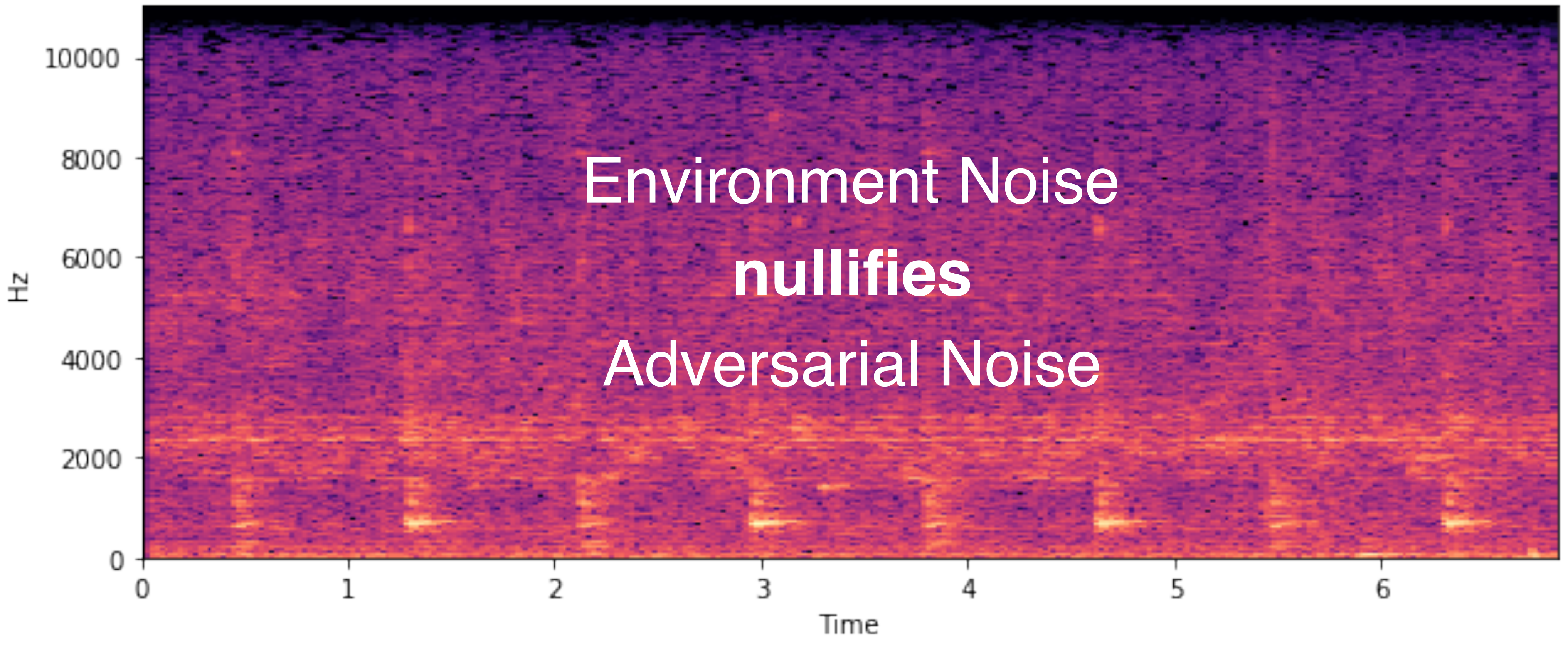


+



Sample adversarial noise

Schönherr et al. [2019]

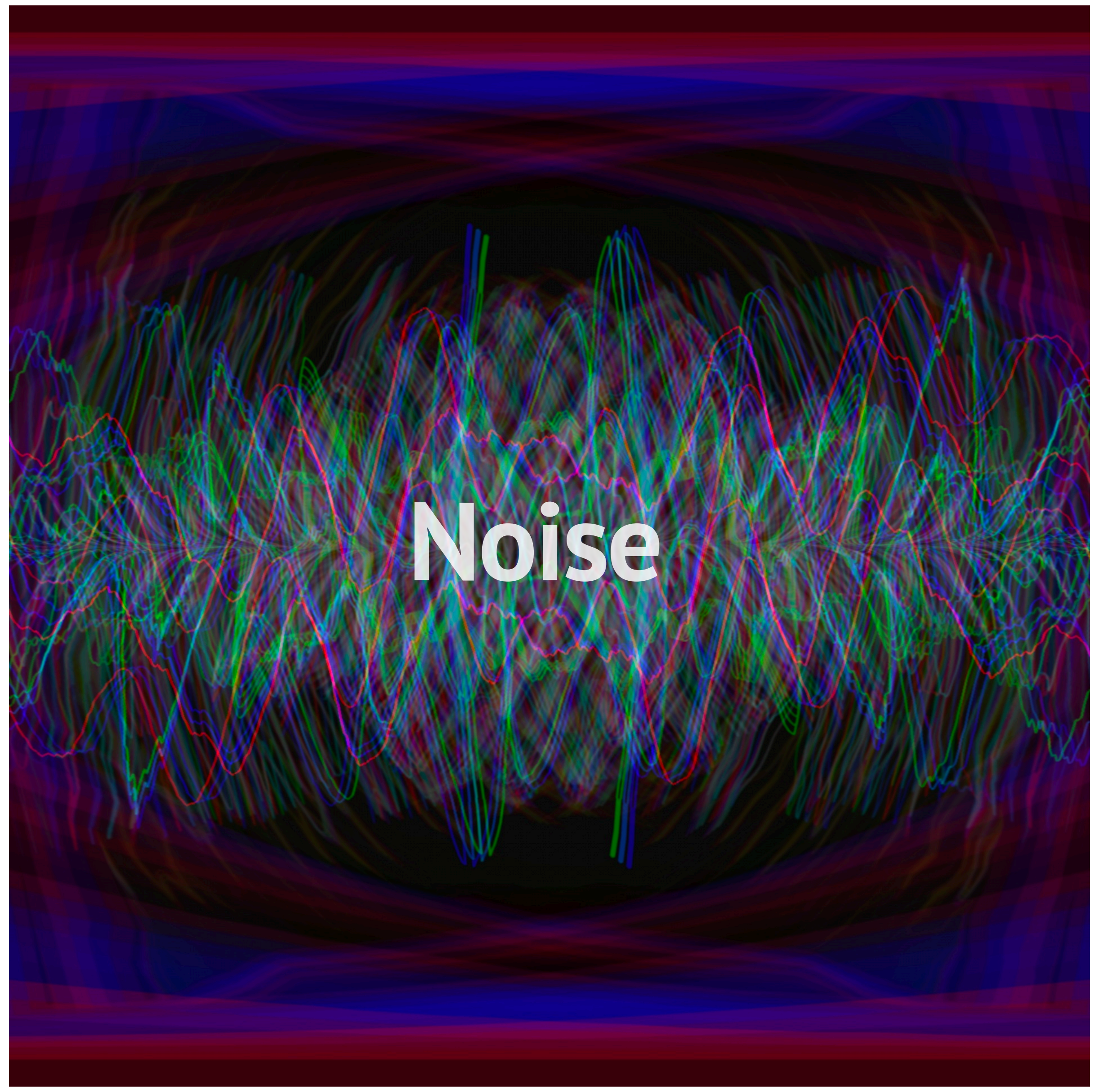


Environment noise at home

Fish tank + clock

Two Big Challenges

The actual Alexa model is a black box



Unstructured adversarial noises are not robust in practice

Contributions

Gray-box over-the-air Denial of Service (Dos) Attack against commercial voice assistant

- A **“gray-box” attack** that leverages the domain transferability of our perturbation. We demonstrated its effect in the real world under **separate audio source** settings.
- A **novel threat model** that allows us to disguise our adversarial attack as **a piece of music with tunable parameters** playable over the air in the physical space.
- Jointly optimizing the attack nature while fitting the threat model to the perturbation achievable by the microphone hearing response of Amazon Alexa. Our **attack budget is very limited** compared with previous works, which makes this challenging.

“Grey Box” Attack

Emulated Wake-word Detect Model

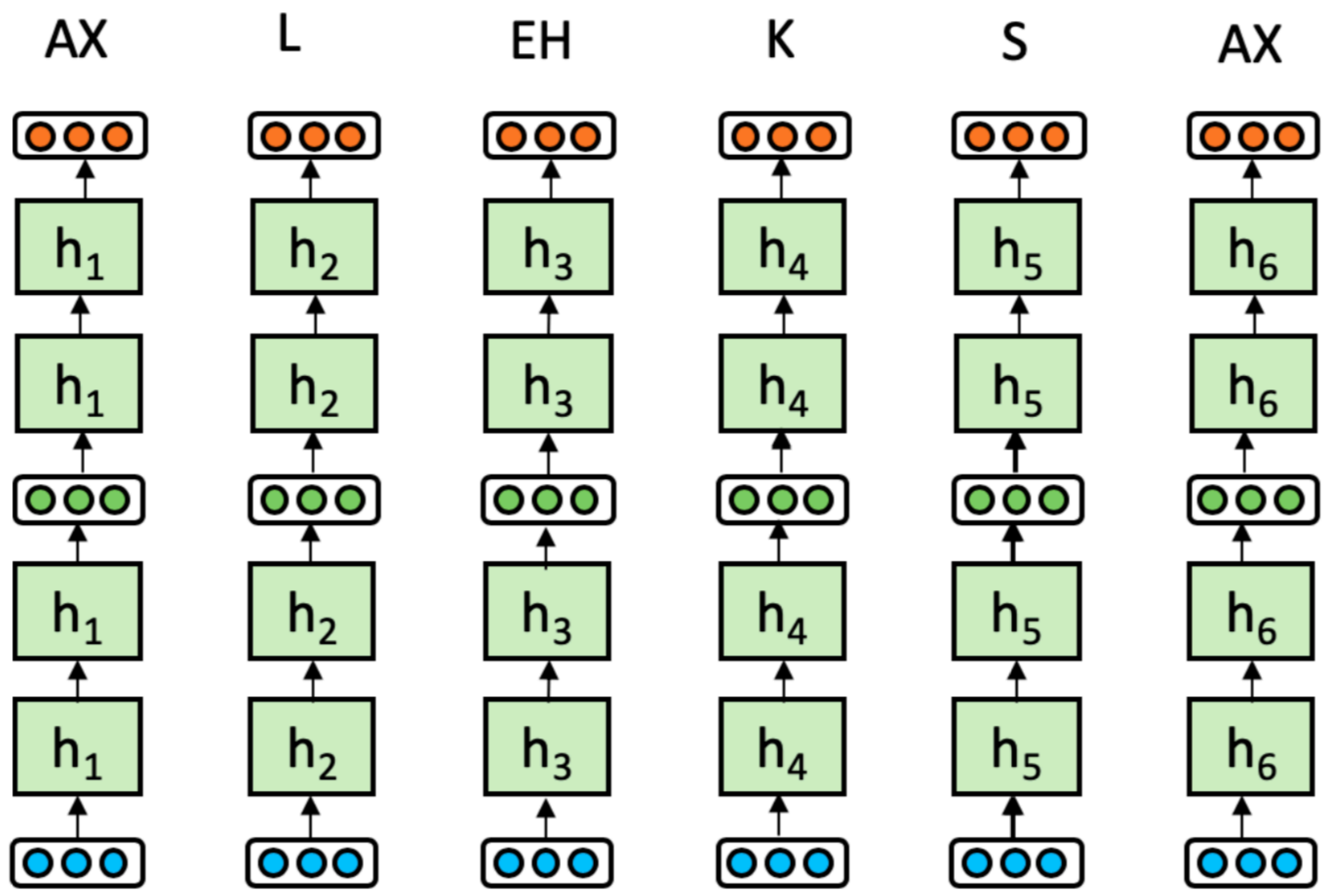


Figure 1: Emulated Model Architecture based on Panchapagesan et al. [2016], Kumatani et al. [2017], Guo et al. [2018]

Detection Error Tradeoff

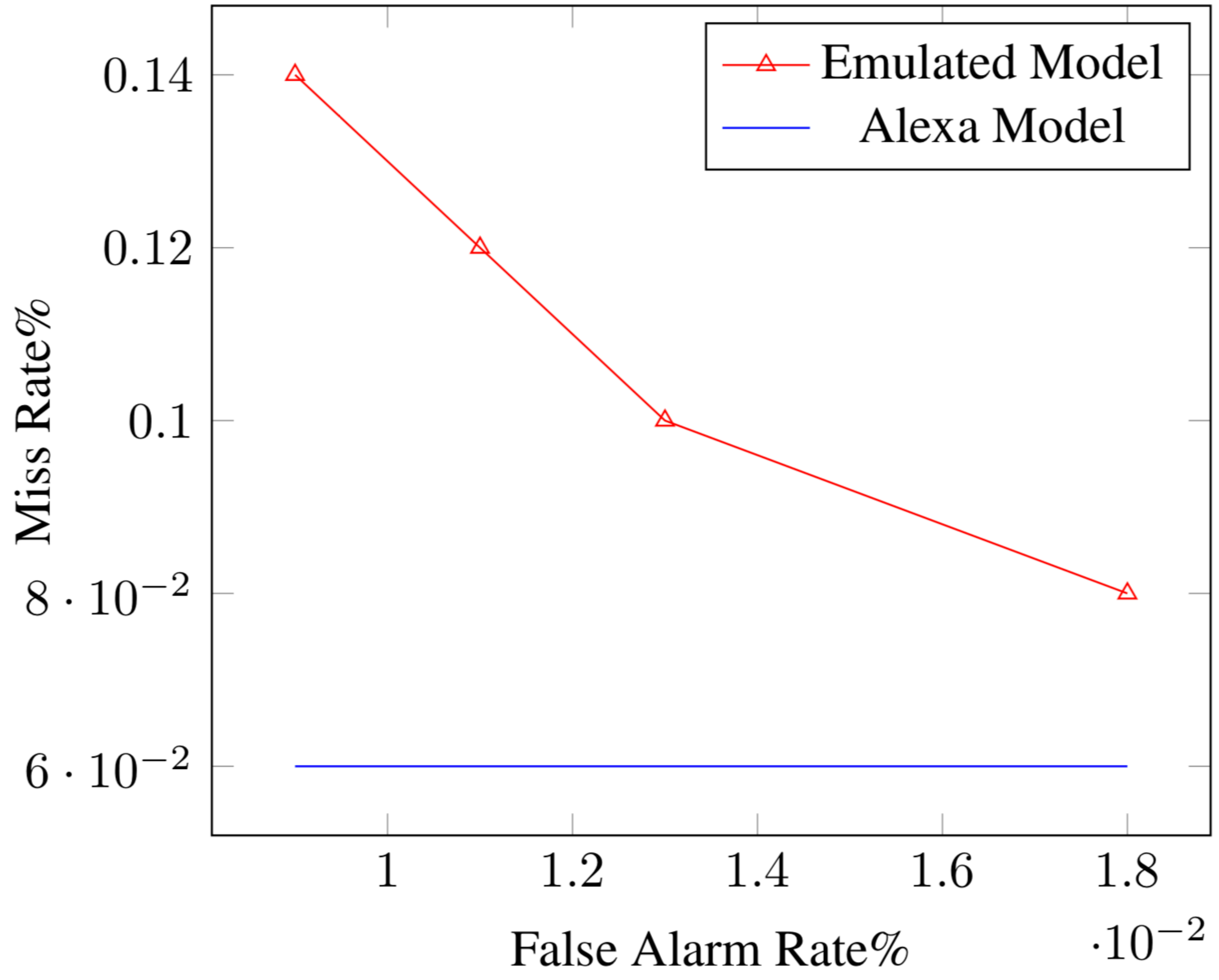


Figure 2: Detection Error Tradeoff Curve. The curve of Alexa model is shown in a flat line as its false alarm rate is not published

Adversarial Music Generation using Physical Modeling Synthesizer

Physical Modeling Synthesizer

Duration

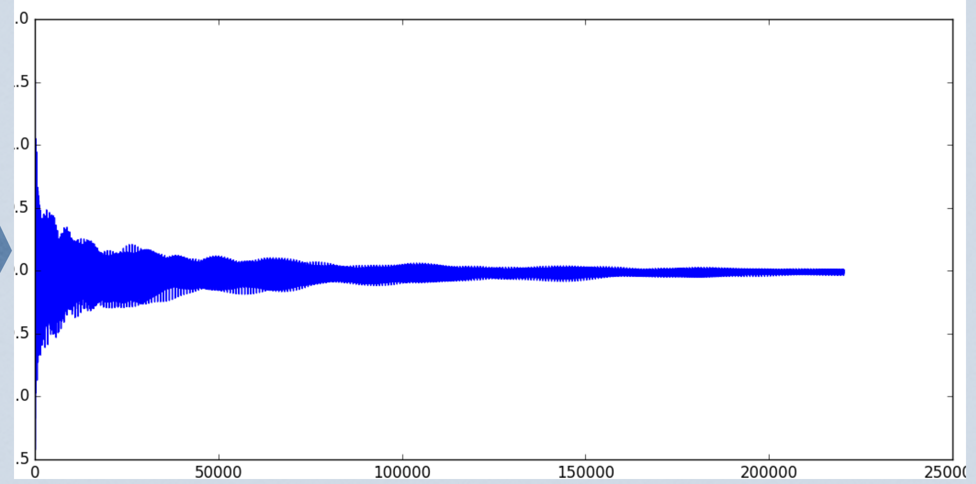
Key

Volume

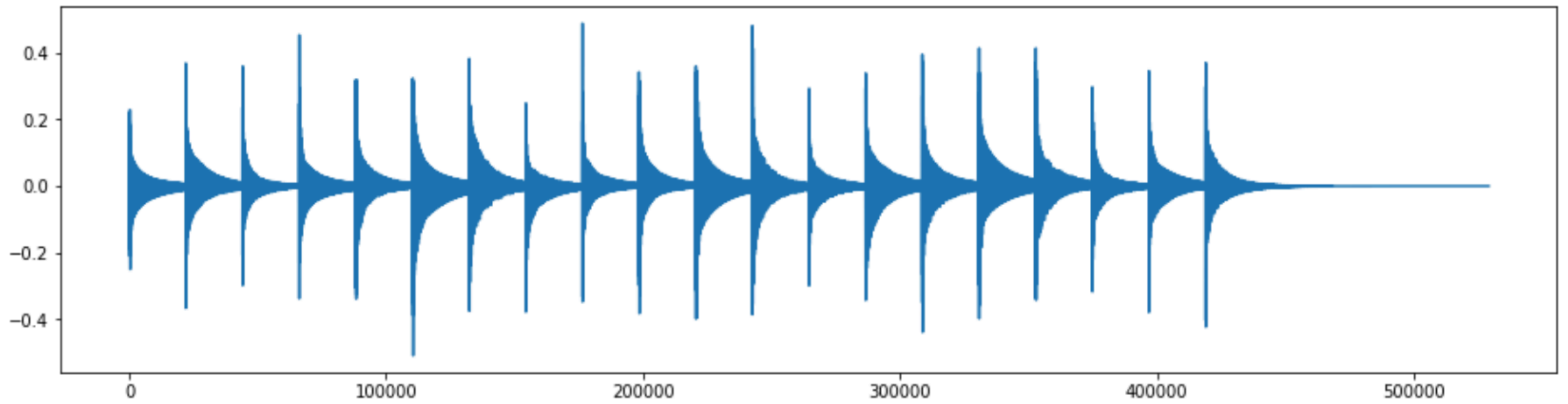
Karplus Strong Generator

Jaffe and Smith [1983]

Iteratively decay average

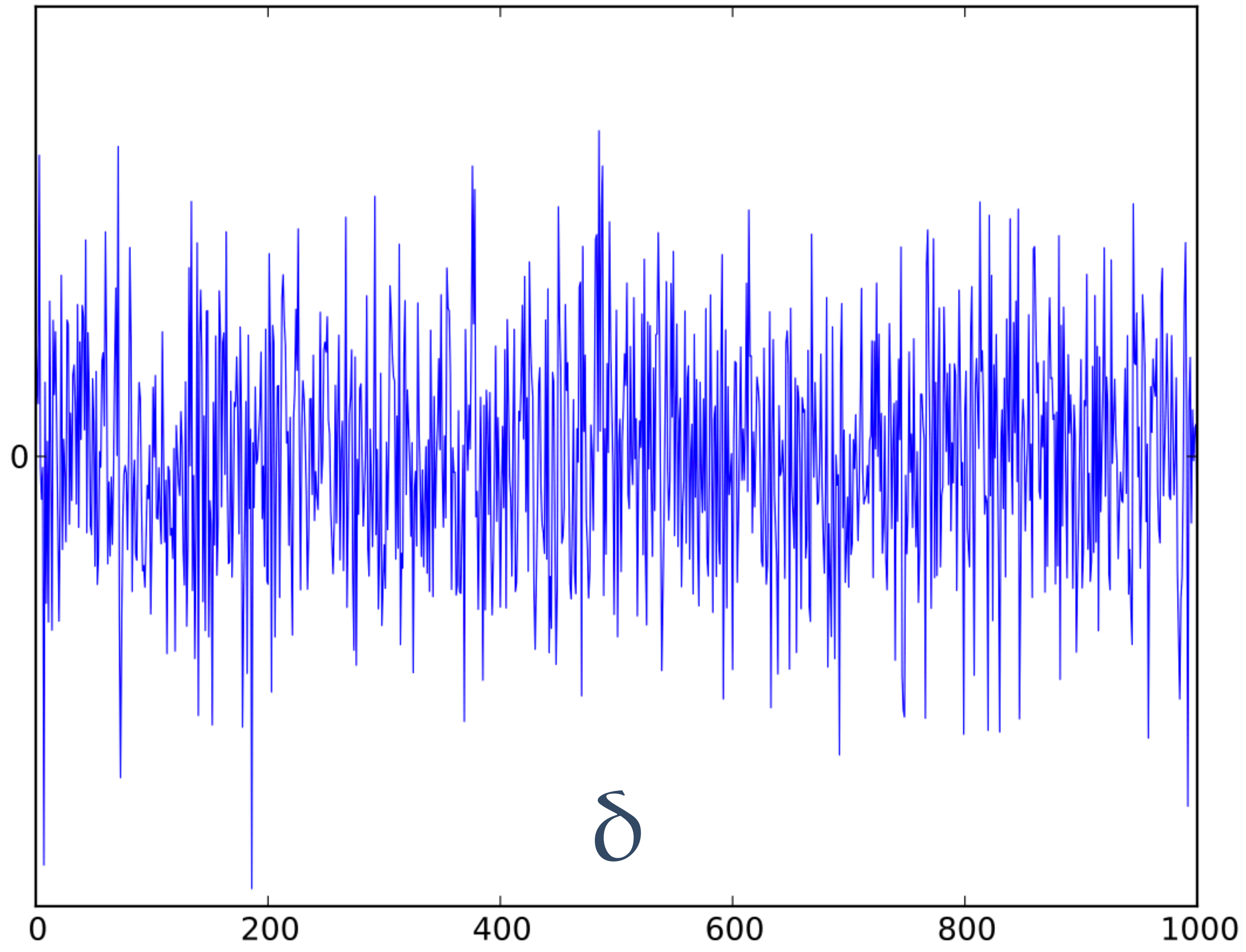


$\theta_{Duration}$ θ_{Key} θ_{Volume}



δ_θ

Unstructured noise



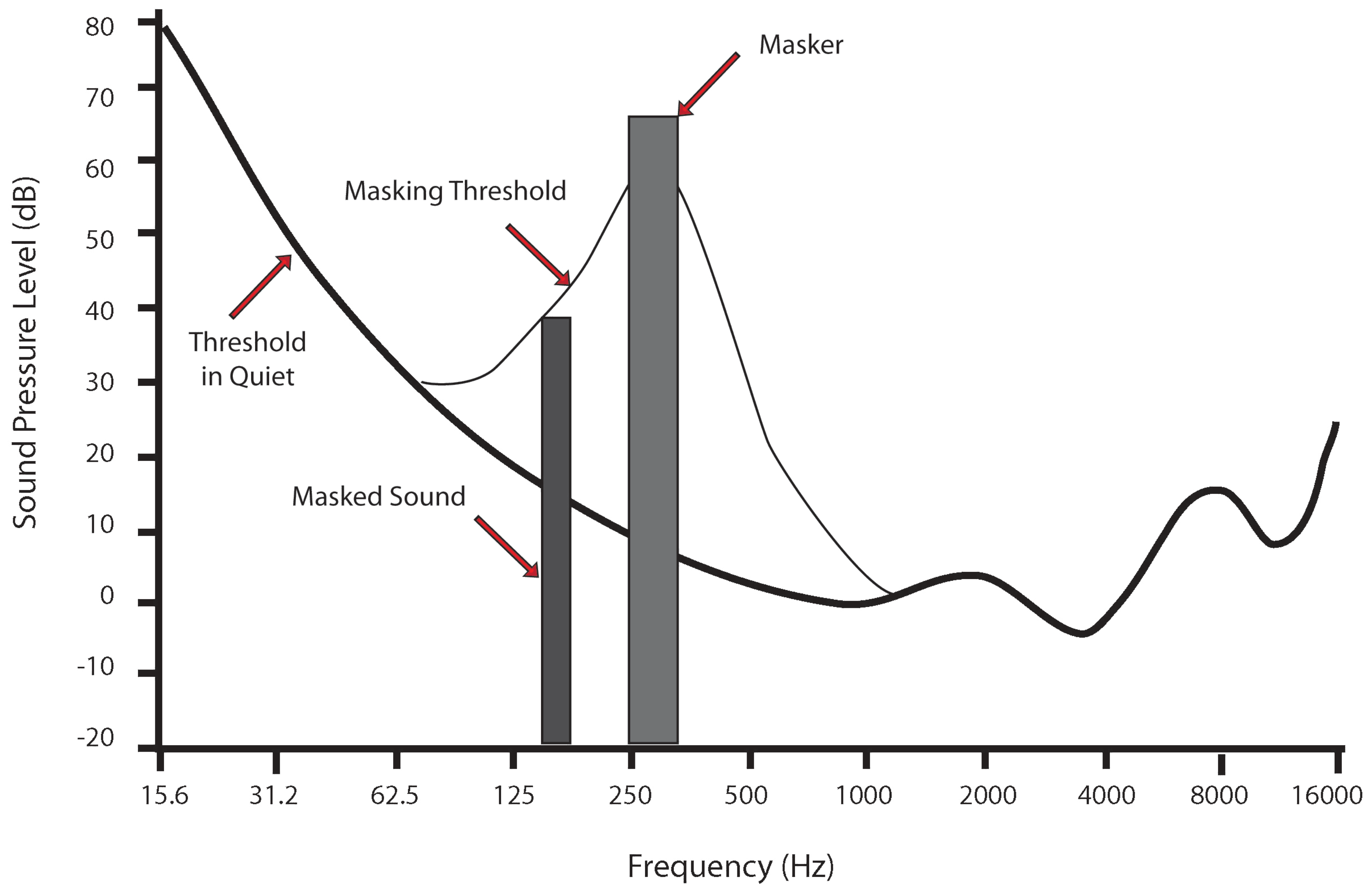
Wikipedia contributors. “White Noise" *Wikipedia*

Juncheng B. Li, Shuhui Qu, Xinjian Li, Joseph Szurley, J. Zico Kolter, Florian Metze

Carnegie Mellon University | BOSCH | Stanford University

Combat Distortion with Limited Attack Budget

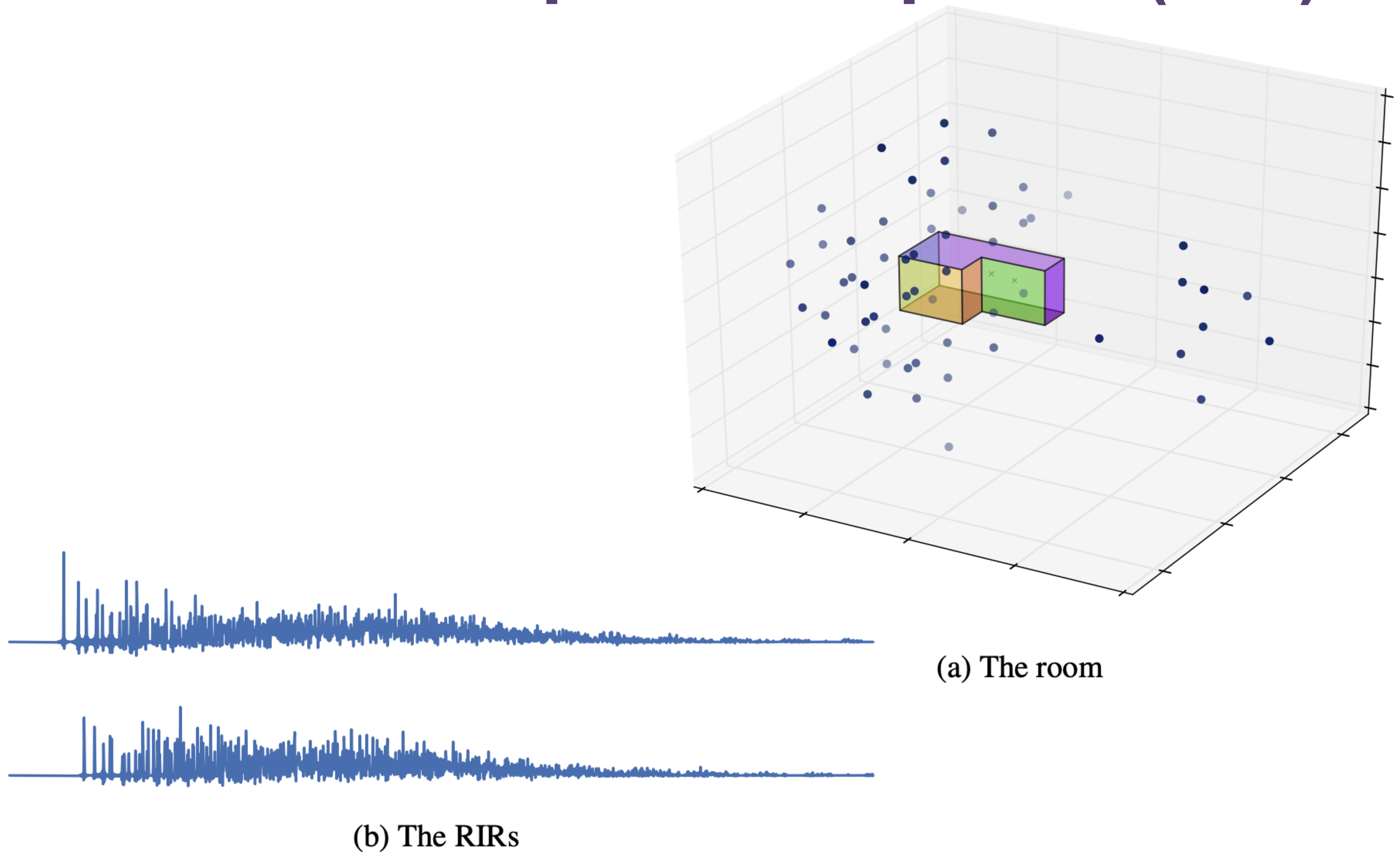
Psychoacoustic Effect



Audio masking graph

Wikipedia contributors. "Psychoacoustics." *Wikipedia*

Room Impulse Response (RIR)



Scheibler et al. [2019],

RIR

Psychoacoustic term

Final Loss: $\max l(x, \delta_\theta, y) = \mathbf{E}_{t\in\mathcal{T},x,y\sim D}[L_{\text{wake}}(f(t(x+\delta_\theta), y)) - \boxed{\alpha \cdot L_\eta(x, \delta_\theta)]}$

 Juncheng B. Li, Shuhui Qu, Xinjian Li, Joseph Szurley, J. Zico Kolter, Florian Metze

Carnegie Mellon University | BOSCH | Stanford University

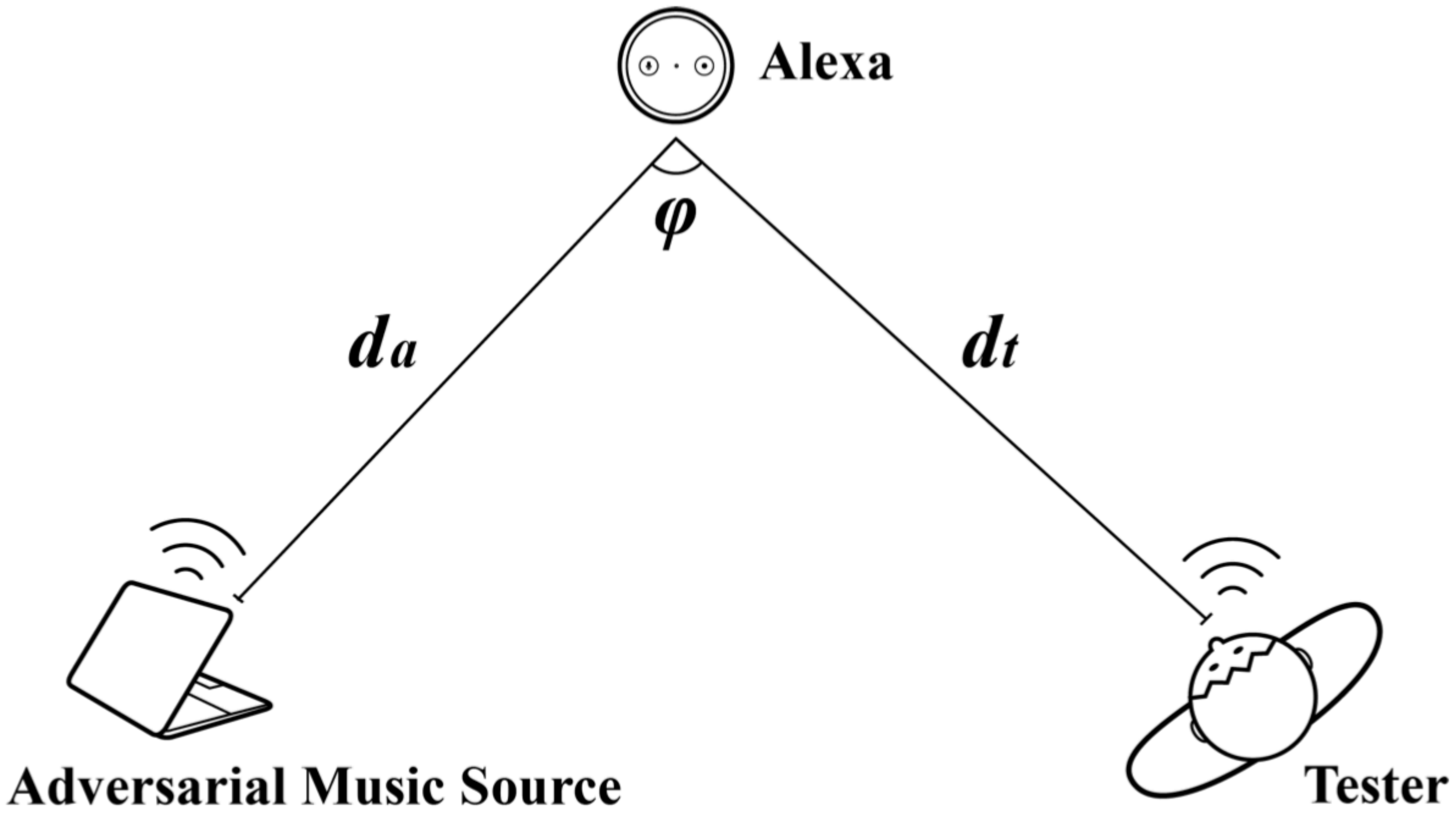
Results

Model	Digital/ Physical	Precision		Recall		F1 Score		# of Sample
		w/o Attack	Attack	w/o Attack	Attack	w/o Attack	Attack	
Emulated Model	Digital	0.97	0.14	0.94	0.11	0.95	0.117	4000
Emulated Model	Physical	0.96	0.12	0.91	0.09	0.934	0.110	100
Alexa	Physical	0.93	0.11	0.92	0.10	0.925	0.110	100

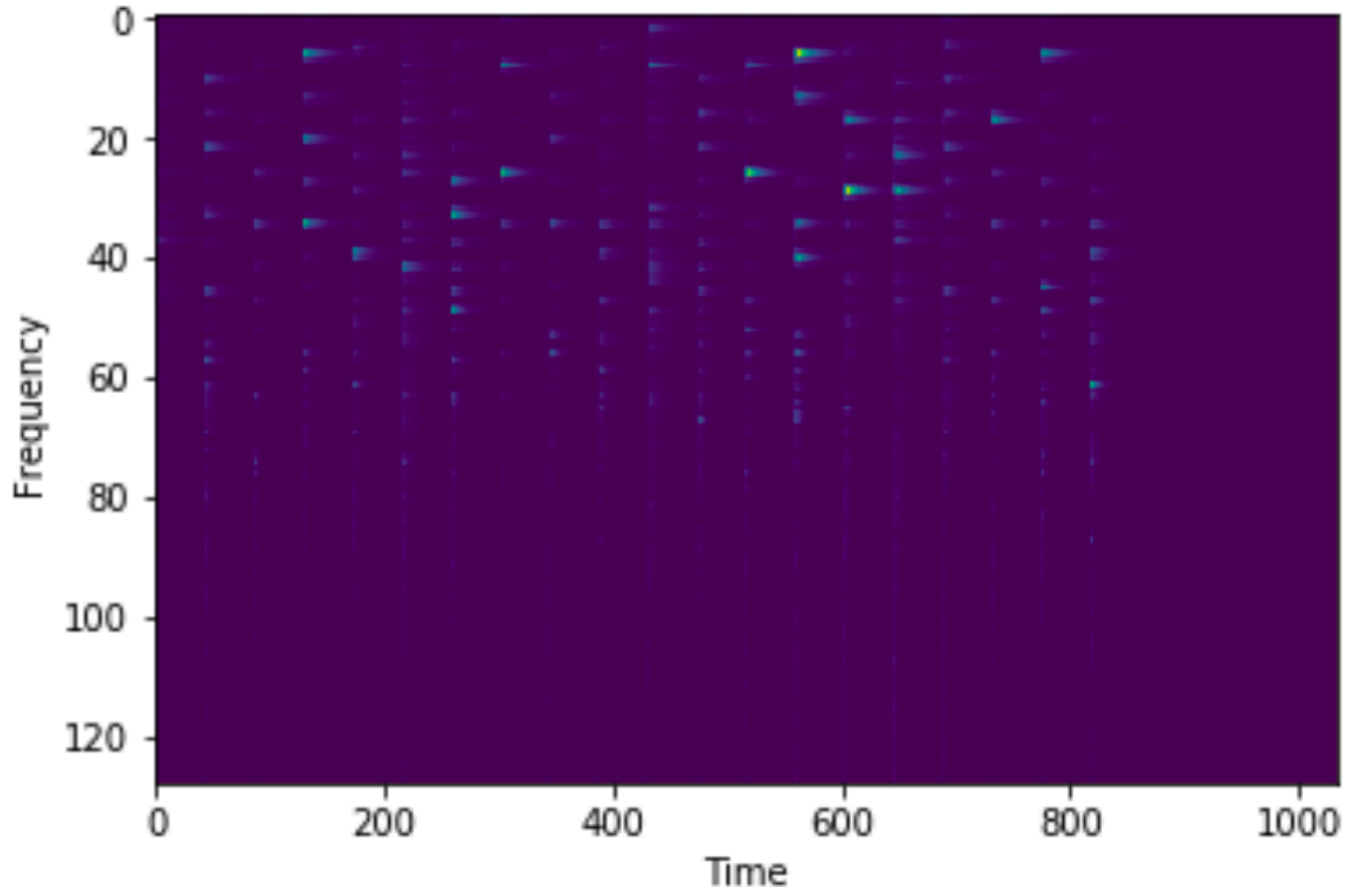
Table 1. Performance of the models with and without attacks in digital and physical testing environments given the number of testing samples

Over-the-air Experiment Setup

Over-the-air testing illustration



Spectrogram of the generated adversarial music



Over-the-air Evaluation

Test Against Alexa		$\phi = 0^\circ$ d_t=			$\phi = 90^\circ$ d_t=			$\phi = 180^\circ$ d_t=		
d_a=	Volume	4.2 ft	7.2 ft	10.2 ft	4.2 ft	7.2 ft	10.2 ft	4.2 ft	7.2 ft	10.2 ft
4.7ft	70 dbA	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10
6.2ft	70 dbA	1/10	0/10	0/10	1/10	0/10	0/10	1/10	2/10	1/10
7.7ft	70 dbA	2/10	0/10	0/10	3/10	1/10	1/10	3/10	3/10	1/10
4.7ft	60 dbA	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10
6.2ft	60 dbA	1/10	1/10	0/10	3/10	1/10	0/10	2/10	2/10	0/10
7.7ft	60 dbA	2/10	1/10	0/10	3/10	2/10	1/10	4/10	3/10	1/10
4.7ft	50 dbA	1/10	2/10	1/10	2/10	2/10	2/10	2/10	2/10	1/10
6.2ft	50 dbA	2/10	3/10	2/10	3/10	3/10	2/10	2/10	3/10	2/10
7.7ft	50 dbA	2/10	3/10	2/10	3/10	2/10	3/10	4/10	3/10	3/10

Carnegie Mellon University | BOSCH | Stanford University


7.2 ft
We tested our Adversarial Music against
Amazon Echo's wake word detection:
"Alexa" in a normal household
environment. In this case, the tester is
standing 7.2ft away from the Amazon Echo

𝄞 *Thank you!*

See you

on Thursday, Dec 12th 10:45-12:45

at East Exhibition Hall B + C #10

Adversarial Music:
Real world audio adversary against wake-word detection systems

Juncheng B. Li, Shuhui Qu, Xinjian Li,
Joseph Szurley, J. Zico Kolter, Florian Metze