CASA

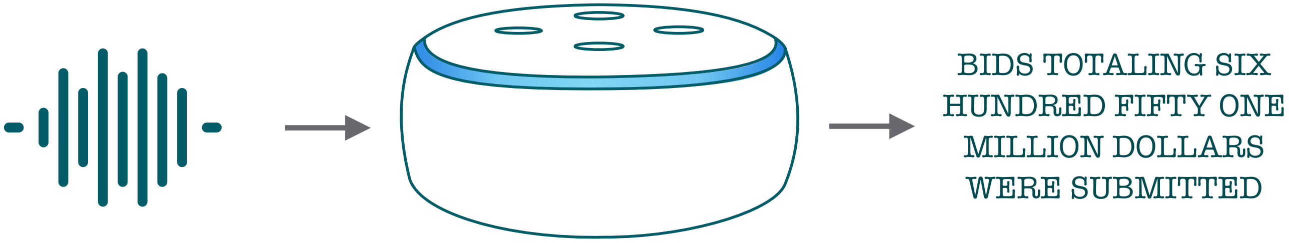
CYBER SECURITY IN THE AGE
OF LARGE-SCALE ADVERSARIES

DOMPTEUR: TAMING AUDIO ADVERSARIAL EXAMPLES

Thorsten Eisenhofer, Lea Schönherr, Joel Frank,
Lars Speckemeier, Dorothea Kolossa, Thorsten Holz

USENIX SECURITY SYMPOSIUM 2021

RUHR
UNIVERSITÄT
BOCHUM
RUB

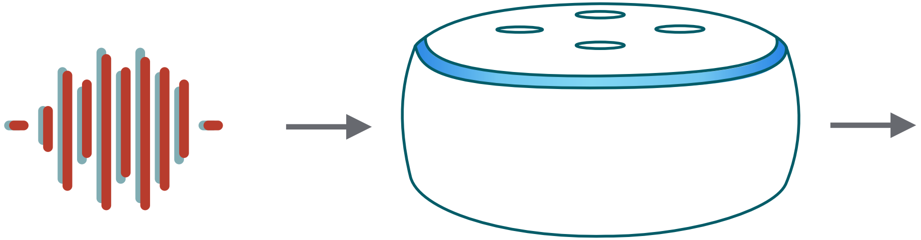


BIDS TOTALING SIX
HUNDRED FIFTY ONE
MILLION DOLLARS
WERE SUBMITTED

Raw Audio Wave

Voice Assistant

Transcription

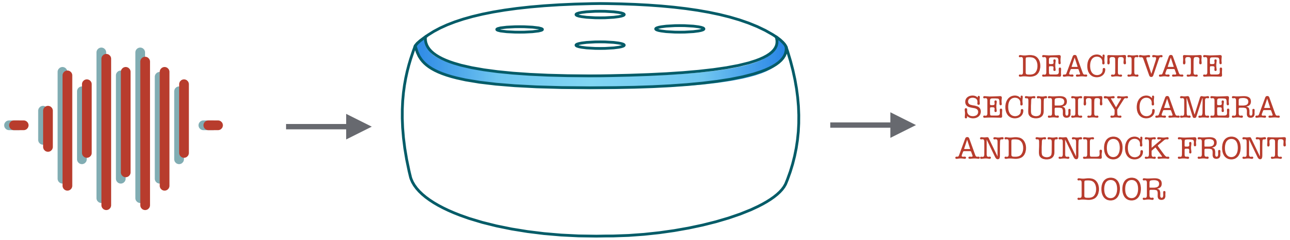


DEACTIVATE SECURITY CAMERA AND UNLOCK FRONT DOOR

Raw Audio Wave

Voice Assistant

Transcription



DEACTIVATE SECURITY CAMERA AND UNLOCK FRONT DOOR

Raw Audio Wave

Voice Assistant

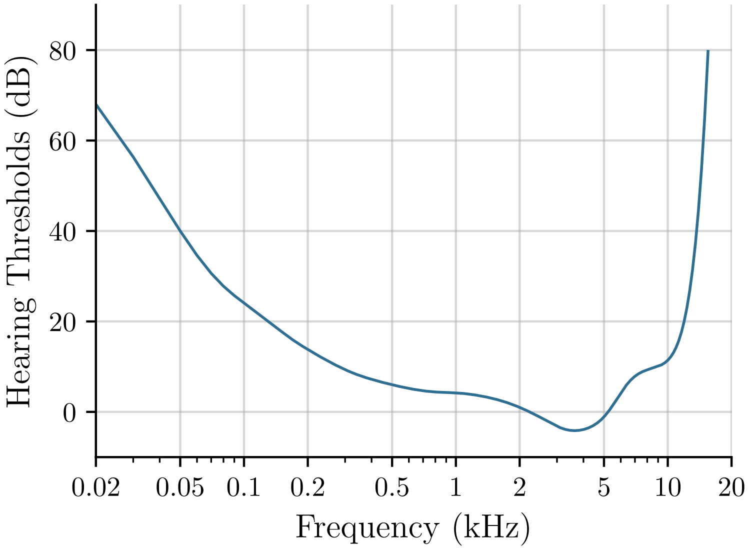
Transcription

*When we **accept** that **adversarial examples** exist, what **else** can we do?*

Gustav Fechner

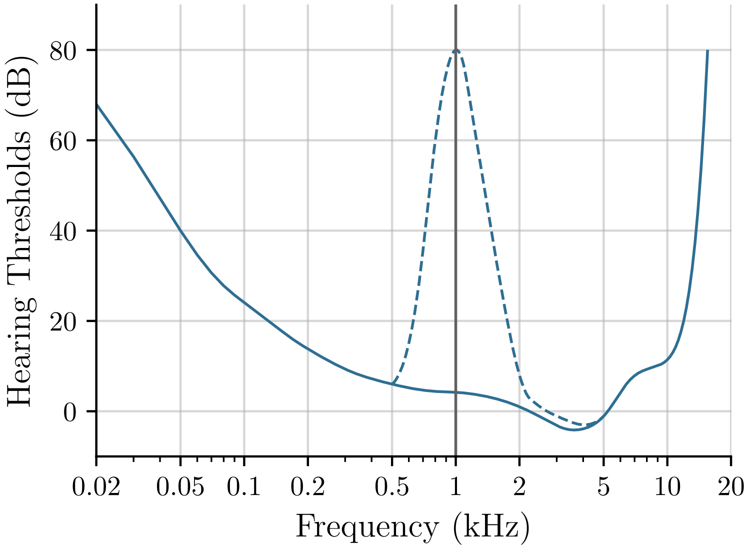
1801-1887

Absolute Hearing Thresholds



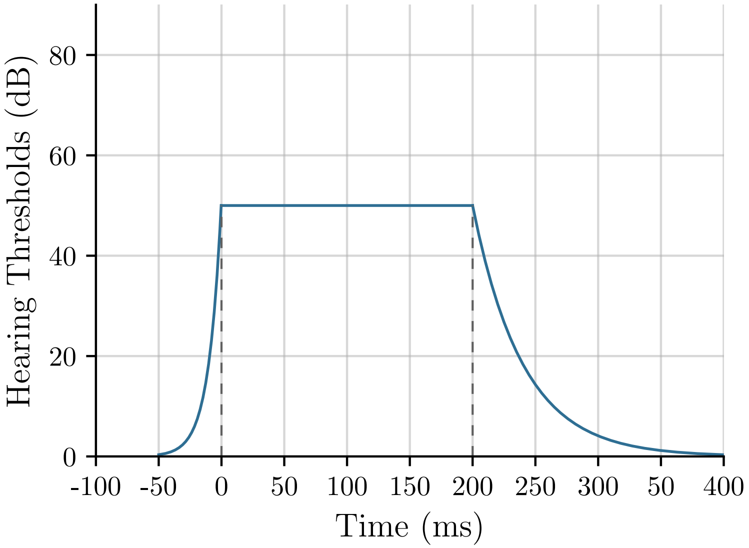
Gustav Fechner
1801-1887

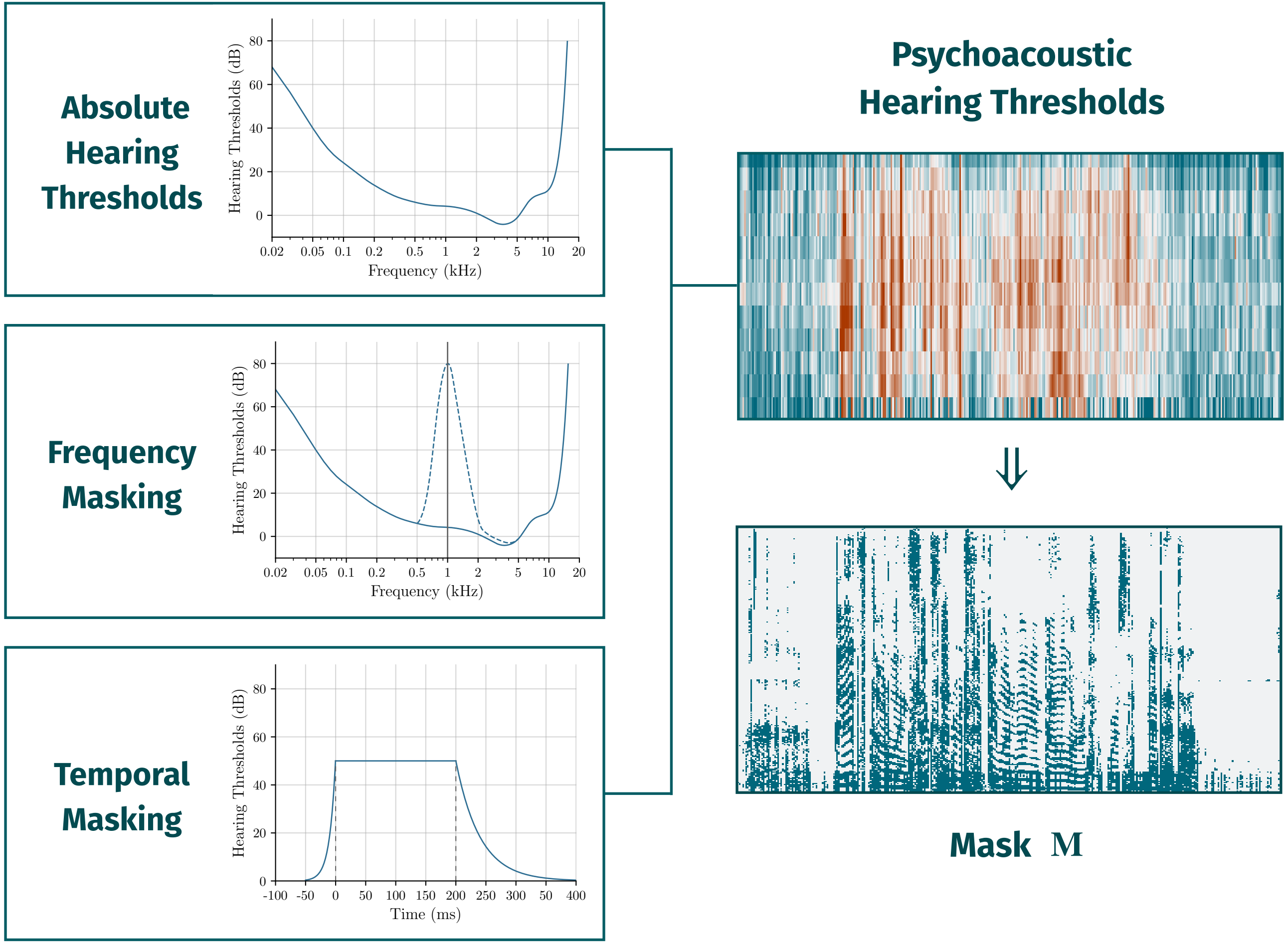
Frequency Masking



Gustav Fechner
1801-1887

Temporal Masking





Psychoacoustic Hearing Thresholds

Absolute Hearing Thresholds

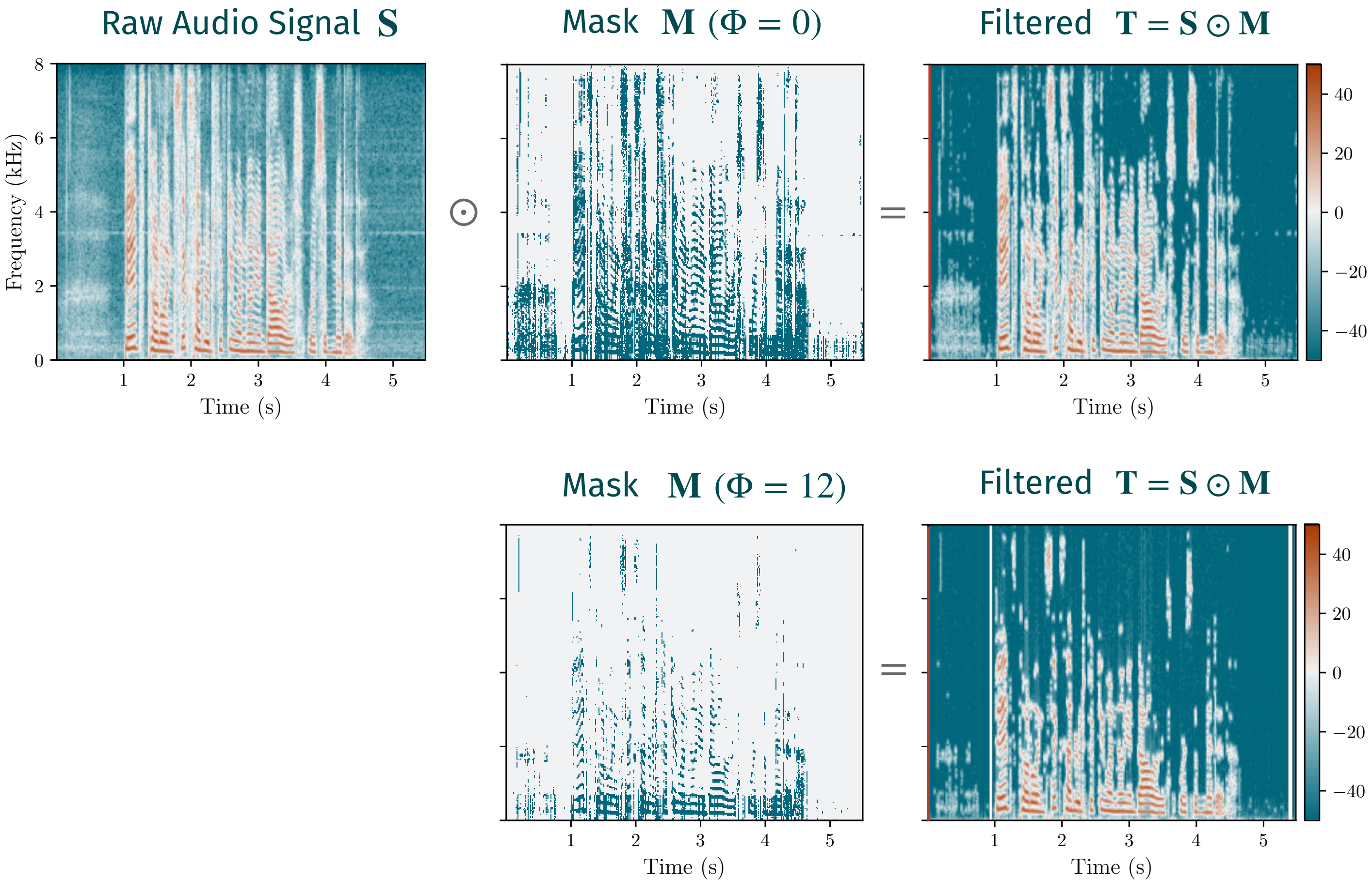
Frequency Masking

Temporal Masking

Mask $\mathbf{M}$

CASA
Cyber Security in the Age of Large-Scale Adversaries

RUHR UNIVERSITÄT BOCHUM RUB



Raw Audio Signal $\mathbf{S}$

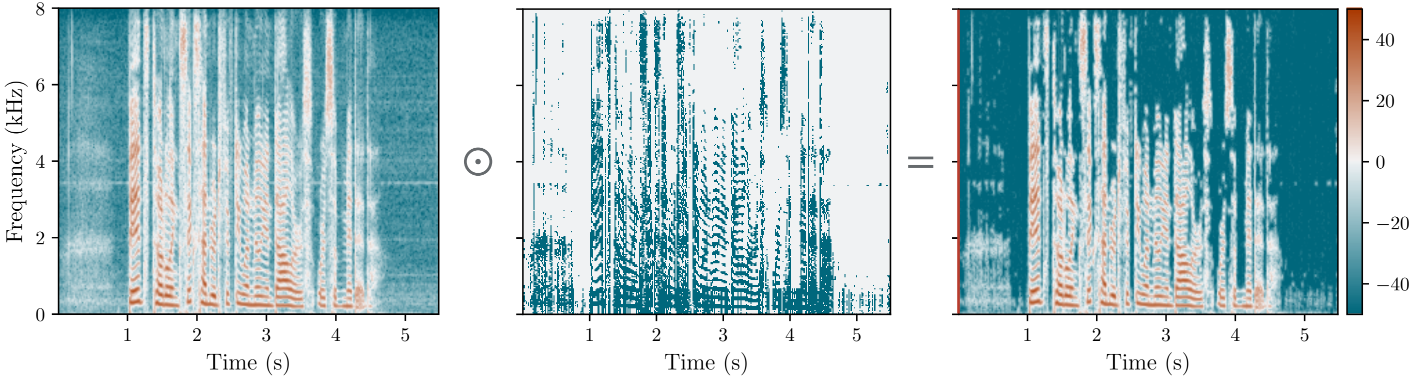
Mask $\mathbf{M}$ $(\Phi = 0)$

Filtered $\mathbf{T} = \mathbf{S} \odot \mathbf{M}$

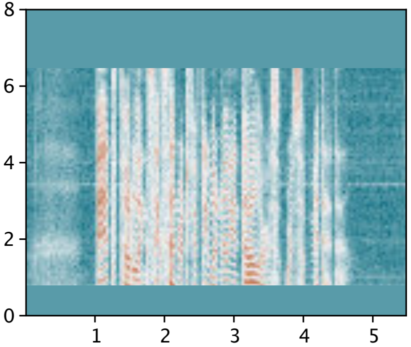
Mask $\mathbf{M}$ $(\Phi = 12)$

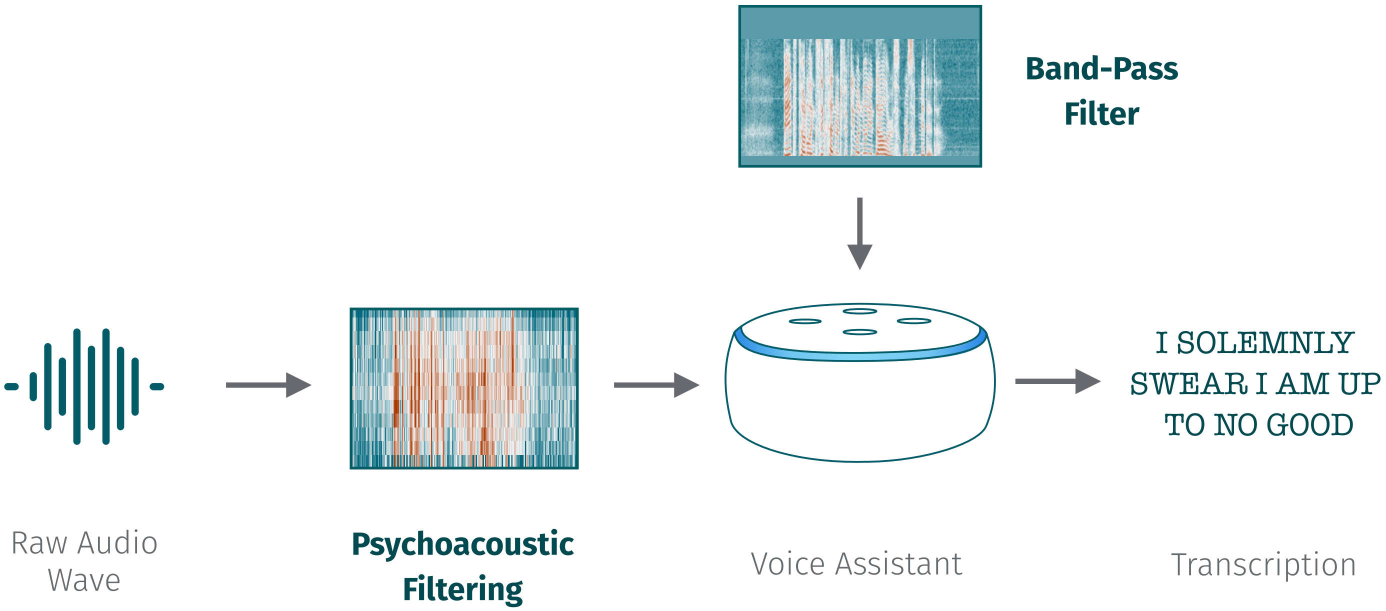
Filtered $\mathbf{T} = \mathbf{S} \odot \mathbf{M}$

Raw Audio Signal $\mathbf{S}$ Mask $\mathbf{M}$ $(\Phi = 0)$ Filtered $\mathbf{T} = \mathbf{S} \odot \mathbf{M}$

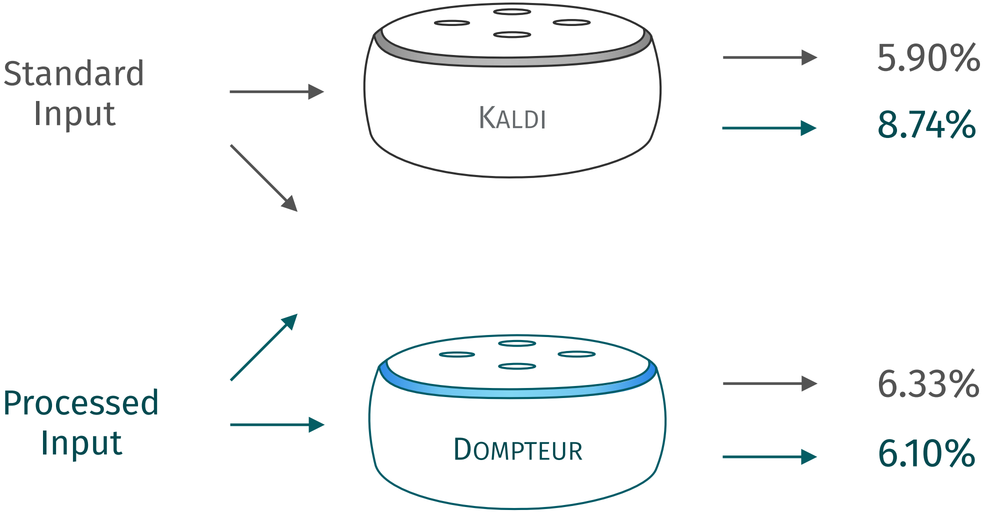


Band-Pass Filter





Implemented DOMPTEUR for **Kaldi** toolkit



Standard Input

Processed Input

KALDI

DOMPTEUR

5.90%

8.74%

6.33%

6.10%

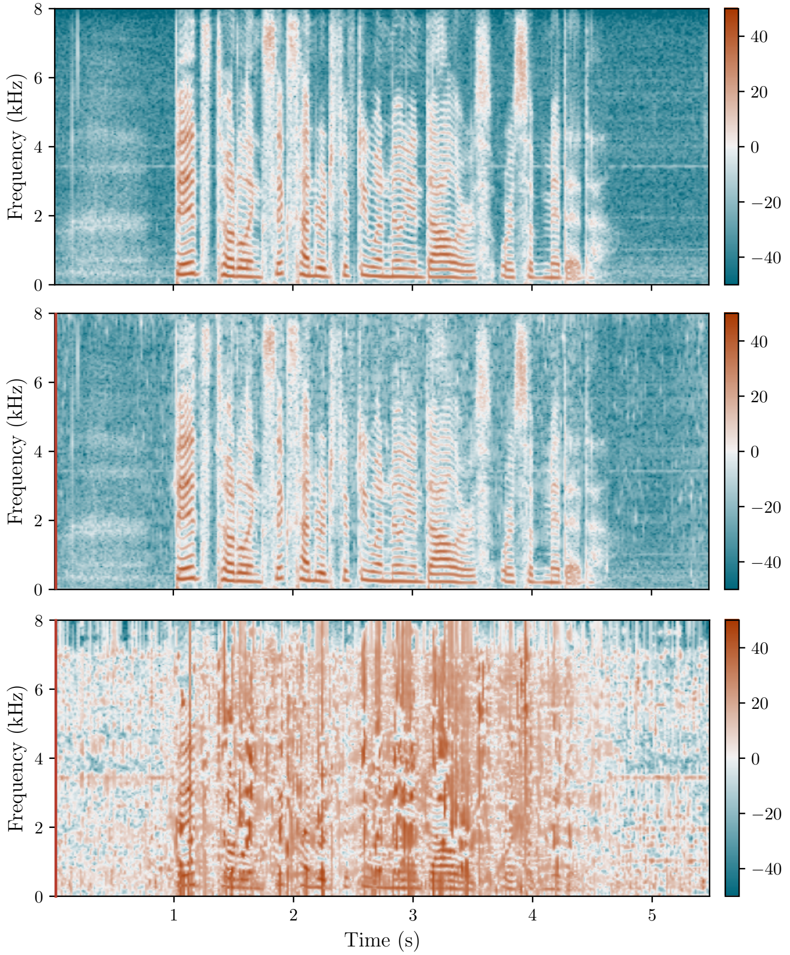
Word Error Rate (WER)

Adversarial Robustness

Strong adaptive, **white-box** attacker

Successful at computing adversarial examples against **Dompteur**

But attack forced into audible ranges and **clearly perceivable**



Unmodified Signal

BIDS TOTALING SIX HUNDRED FIFTY ONE MILLION DOLLARS WERE SUBMITTED

Kaldi

SEND SECRET FINANCIAL REPORT

Dompteur $\Phi = 12$

SEND SECRET FINANCIAL REPORT

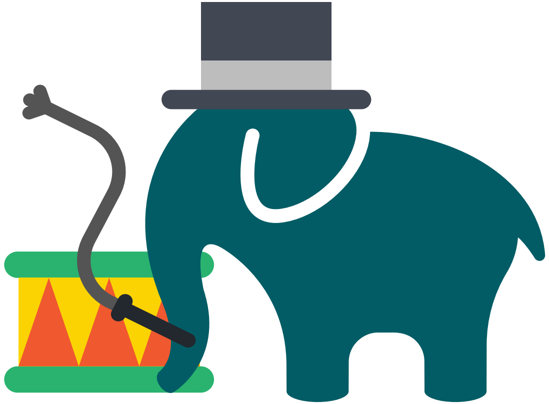
TAKEAWAYS

Adversarial examples seem to be **inevitable**

New Perspective: Make attack **noticeable**

Psychoacoustics effective to force attack into **audible ranges**

Code, Examples and Models available at
github.com/rub-syssec/dompteur



THANK YOU!