DOMPETEUR: TAMING AUDIO ADVERSARIAL EXAMPLES

Thorsten Eisenhofer, Lea Schönherr, Joel Frank, Lars Speckemeier, Dorothea Kolossa, Thorsten Holz
BIDS TOTALING SIX HUNDRED FIFTY ONE MILLION DOLLARS WERE SUBMITTED
DEACTIVATE SECURITY CAMERA AND UNLOCK FRONT DOOR
When we **accept** that **adversarial examples** exist, what **else** can we do?
Gustav Fechner
1801-1887

Absolute Hearing Thresholds

Hearing Thresholds (dB)

Frequency (kHz)

0 20 40 60 80
0.02 0.05 0.1 0.2 0.5 1 2 5 10 20
Frequency Masking

Hearing Thresholds (dB)

Frequency (kHz)
Temporal Masking

Hearing Thresholds (dB)

Time (ms)
Psychoacoustic Hearing Thresholds

Absolute Hearing Thresholds

Frequency Masking

Temporal Masking

Mask M
Raw Audio Signal $S$

Mask $M$ ($\Phi = 0$)

Filtered $T = S \odot M$

Band-Pass Filter
Raw Audio
Wave

Psychoacoustic
Filtering

Band-Pass
Filter

I SOLEMNLY
SWEAR I AM UP
TO NO GOOD

Voice Assistant

Transcription

Implemented DOMPETEUR for Kaldi toolkit
Word Error Rate (WER)

Standard Input:
- KALDI
  - Input: 5.90%
  - Processed: 8.74%

Processed Input:
- DOMPETEUR
  - Input: 6.33%
  - Processed: 6.10%
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

DOMPETEUR $\Phi = 12$

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!