Adversarially Robust Speech Recognition

Thorsten Eisenhofer and Lea Schönherr

CASA
Cyber Security in the Age of Large-Scale Adversaries

Ruhr Universität Bochum
RUB

Large–Scale
Adversary

HUB C
Intelligent Security Systems

Adversarial Examples and Defenses for Automatic Speech Recognition Systems

Dorothea Kolossa

Cognitive Signal Processing

Lea Schönherr

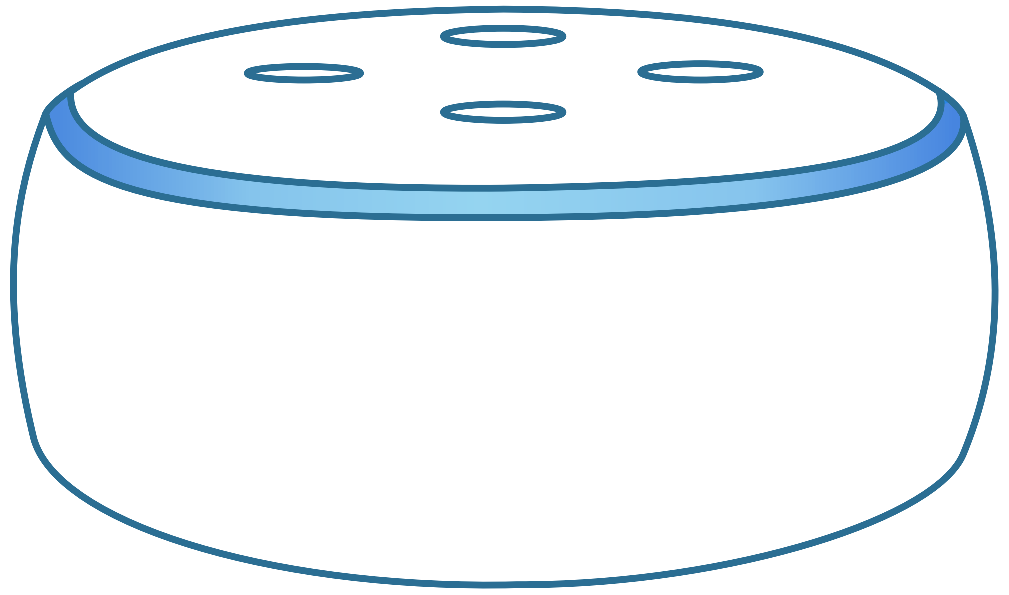
Postdoctoral Researcher

Thorsten Holz

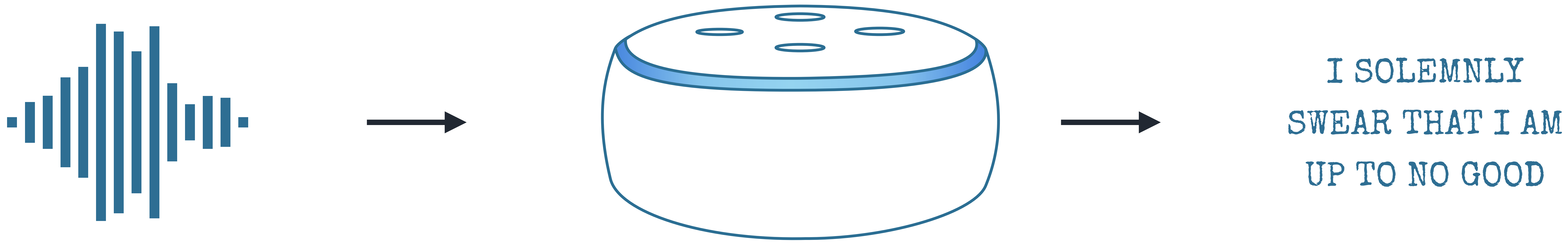
Systems Security

Thorsten Eisenhofer

3rd year PhD Student



Voice Assistant

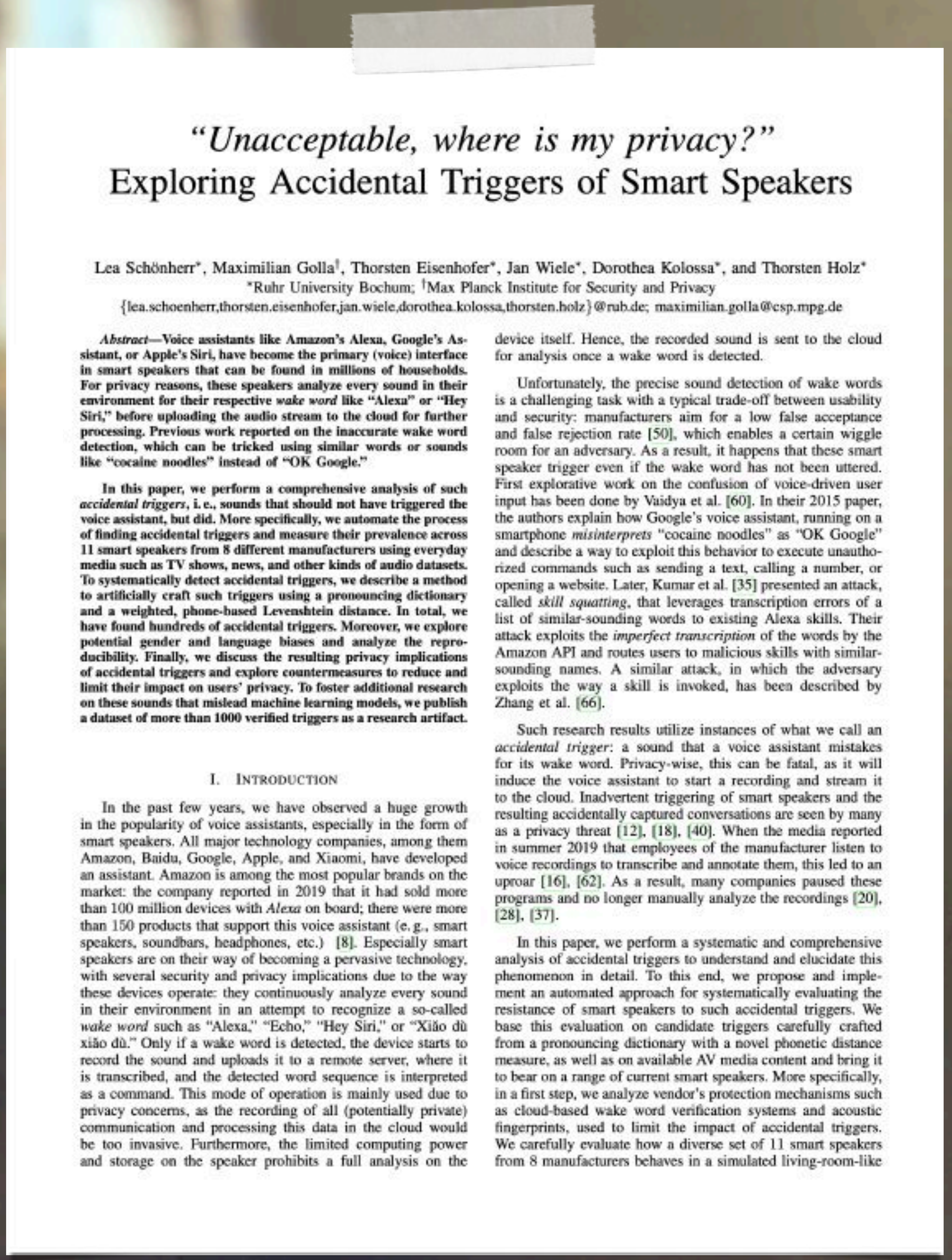


I SOLEMNLY
SWEAR THAT I AM
UP TO NO GOOD

Raw Audio Wave

Voice Assistant

Transcription



"Unacceptable, where is my privacy?" Exploring Accidental Triggers of Smart Speakers

Lea Schönherr*, Maximilian Golla†, Thorsten Eisenhofer*, Jan Wiele*, Dorothea Kolossa*, and Thorsten Holz*
*Ruhr University Bochum; †Max Planck Institute for Security and Privacy
{lea.schoenherr,thorsten.eisenhofer,jan.wiele,dorothea.kolossa,thorsten.holz}@rub.de; maximilian.golla@csp.mpg.de

***Abstract*—Voice assistants like Amazon's Alexa, Google's Assistant, or Apple's Siri, have become the primary (voice) interface in smart speakers that can be found in millions of households. For privacy reasons, these speakers analyze every sound in their environment for their respective *wake word* like "Alexa" or "Hey Siri," before uploading the audio stream to the cloud for further processing. Previous work reported on the inaccurate wake word detection, which can be tricked using similar words or sounds like "cocaine noodles" instead of "OK Google."**

In this paper, we perform a comprehensive analysis of such *accidental triggers*, i. e., sounds that should not have triggered the voice assistant, but did. More specifically, we automate the process of finding accidental triggers and measure their prevalence across 11 smart speakers from 8 different manufacturers using everyday media such as TV shows, news, and other kinds of audio datasets. To systematically detect accidental triggers, we describe a method to artificially craft such triggers using a pronouncing dictionary and a weighted, phone-based Levenshtein distance. In total, we have found hundreds of accidental triggers. Moreover, we explore potential gender and language biases and analyze the reproducibility. Finally, we discuss the resulting privacy implications of accidental triggers and explore countermeasures to reduce and limit their impact on users' privacy. To foster additional research on these sounds that mislead machine learning models, we publish a dataset of more than 1000 verified triggers as a research artifact.

I. Introduction

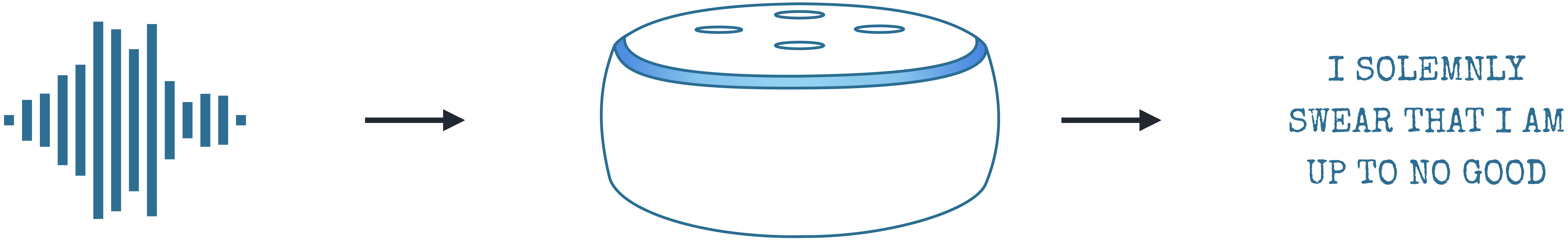
In the past few years, we have observed a huge growth in the popularity of voice assistants, especially in the form of smart speakers. All major technology companies, among them Amazon, Baidu, Google, Apple, and Xiaomi, have developed an assistant. Amazon is among the most popular brands on the market: the company reported in 2019 that it had sold more than 100 million devices with *Alexa* on board; there were more than 150 products that support this voice assistant (e. g., smart speakers, soundbars, headphones, etc.) [8]. Especially smart speakers are on their way of becoming a pervasive technology, with several security and privacy implications due to the way these devices operate: they continuously analyze every sound in their environment in an attempt to recognize a so-called *wake word* such as "Alexa," "Echo," "Hey Siri," or "Xiǎo dù xiǎo dù." Only if a wake word is detected, the device starts to record the sound and uploads it to a remote server, where it is transcribed, and the detected word sequence is interpreted as a command. This mode of operation is mainly used due to privacy concerns, as the recording of all (potentially private) communication and processing this data in the cloud would be too invasive. Furthermore, the limited computing power and storage on the speaker prohibits a full analysis on the device itself. Hence, the recorded sound is sent to the cloud for analysis once a wake word is detected.

Unfortunately, the precise sound detection of wake words is a challenging task with a typical trade-off between usability and security: manufacturers aim for a low false acceptance and false rejection rate [50], which enables a certain wiggle room for an adversary. As a result, it happens that these smart speaker trigger even if the wake word has not been uttered. First explorative work on the confusion of voice-driven user input has been done by Vaidya et al. [60]. In their 2015 paper, the authors explain how Google's voice assistant, running on a smartphone *misinterprets* "cocaine noodles" as "OK Google" and describe a way to exploit this behavior to execute unauthorized commands such as sending a text, calling a number, or opening a website. Later, Kumar et al. [35] presented an attack, called *skill squatting*, that leverages transcription errors of a list of similar-sounding words to existing Alexa skills. Their attack exploits the *imperfect transcription* of the words by the Amazon API and routes users to malicious skills with similar-sounding names. A similar attack, in which the adversary exploits the way a skill is invoked, has been described by Zhang et al. [66].

Such research results utilize instances of what we call an *accidental trigger*: a sound that a voice assistant mistakes for its wake word. Privacy-wise, this can be fatal, as it will induce the voice assistant to start a recording and stream it to the cloud. Inadvertent triggering of smart speakers and the resulting accidentally captured conversations are seen by many as a privacy threat [12], [18], [40]. When the media reported in summer 2019 that employees of the manufacturer listen to voice recordings to transcribe and annotate them, this led to an uproar [16], [62]. As a result, many companies paused these programs and no longer manually analyze the recordings [20], [28], [37].

In this paper, we perform a systematic and comprehensive analysis of accidental triggers to understand and elucidate this phenomenon in detail. To this end, we propose and implement an automated approach for systematically evaluating the resistance of smart speakers to such accidental triggers. We base this evaluation on candidate triggers carefully crafted from a pronouncing dictionary with a novel phonetic distance measure, as well as on available AV media content and bring it to bear on a range of current smart speakers. More specifically, in a first step, we analyze vendor's protection mechanisms such as cloud-based wake word verification systems and acoustic fingerprints, used to limit the impact of accidental triggers. We carefully evaluate how a diverse set of 11 smart speakers from 8 manufacturers behaves in a simulated living-room-like

In Submission @
Computer Speech and Language



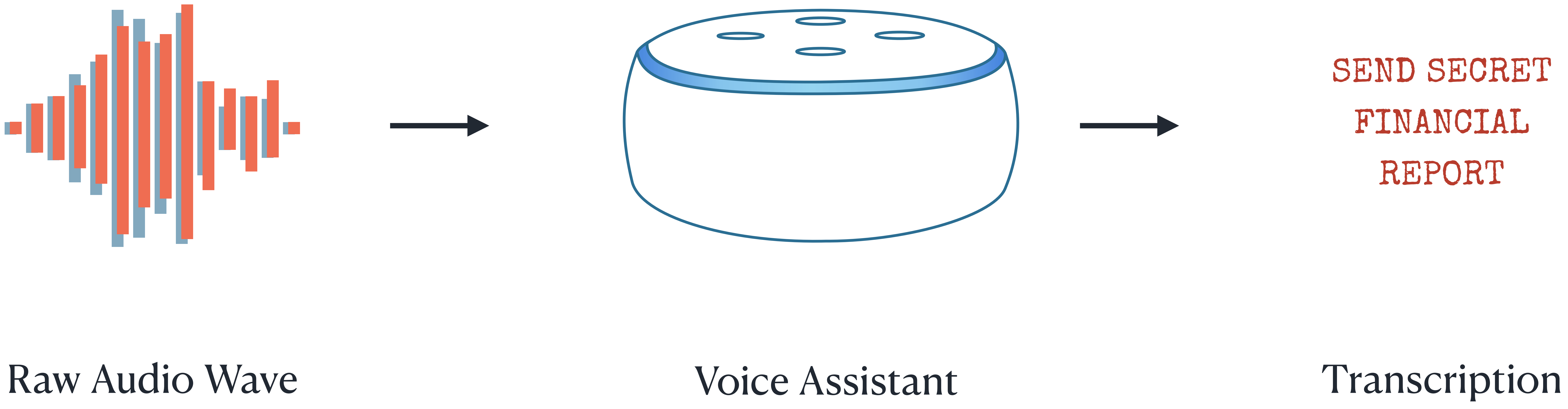
I SOLEMNLY SWEAR THAT I AM UP TO NO GOOD

Raw Audio Wave

Voice Assistant

Transcription

What happens if we add an active attacker to this scenario?



SEND SECRET FINANCIAL REPORT

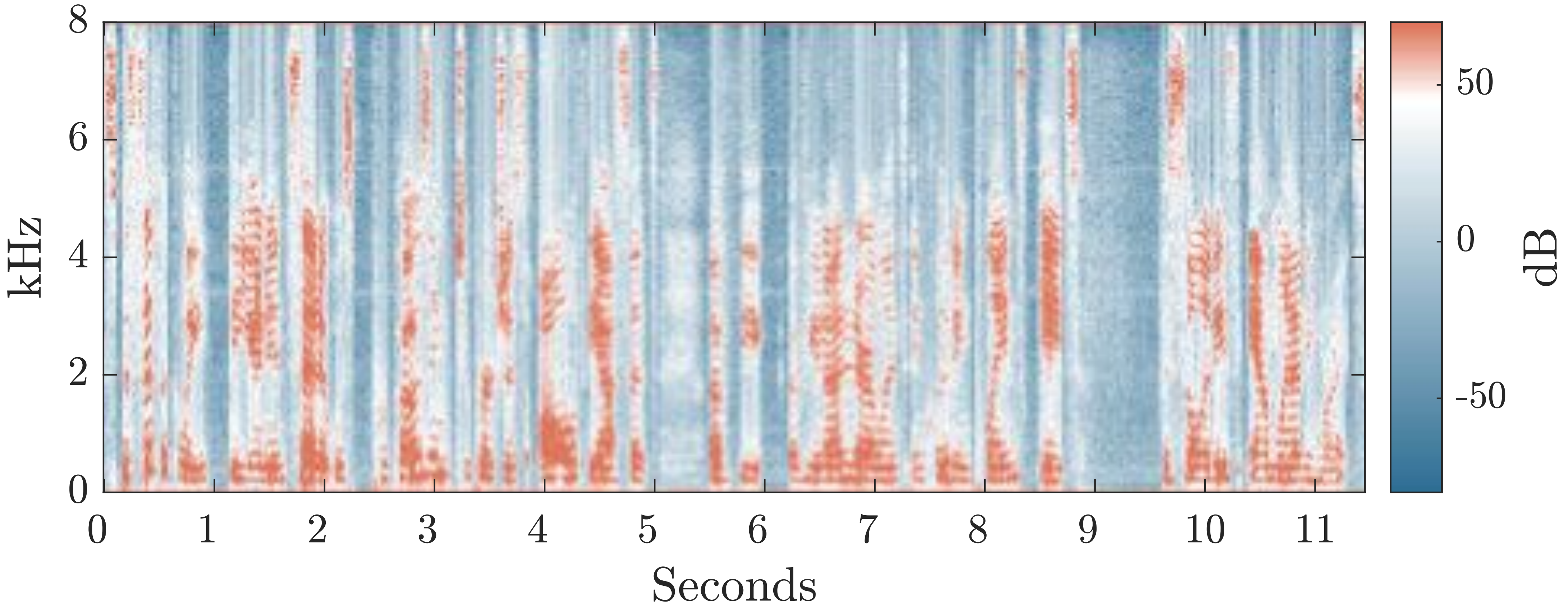
Raw Audio Wave

Voice Assistant

Transcription

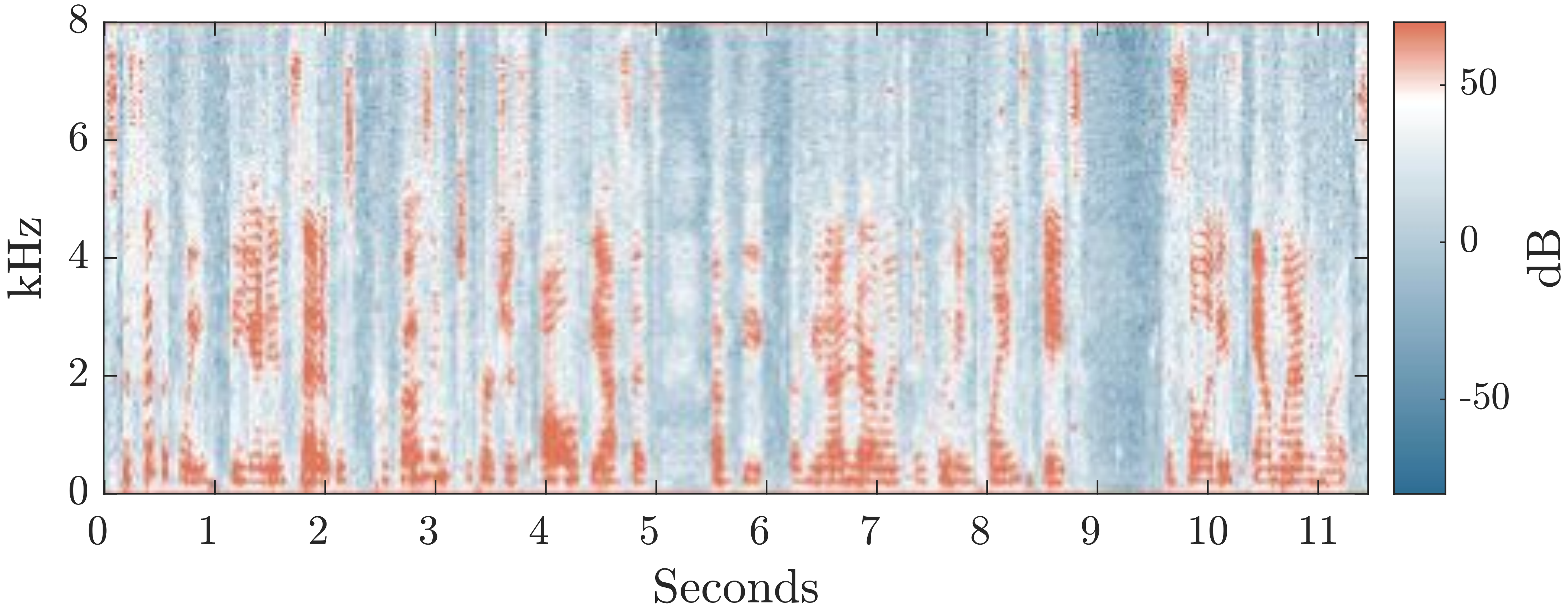
Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, Dorothea Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," Network and Distributed System Security Symposium (NDSS), 2019

Original

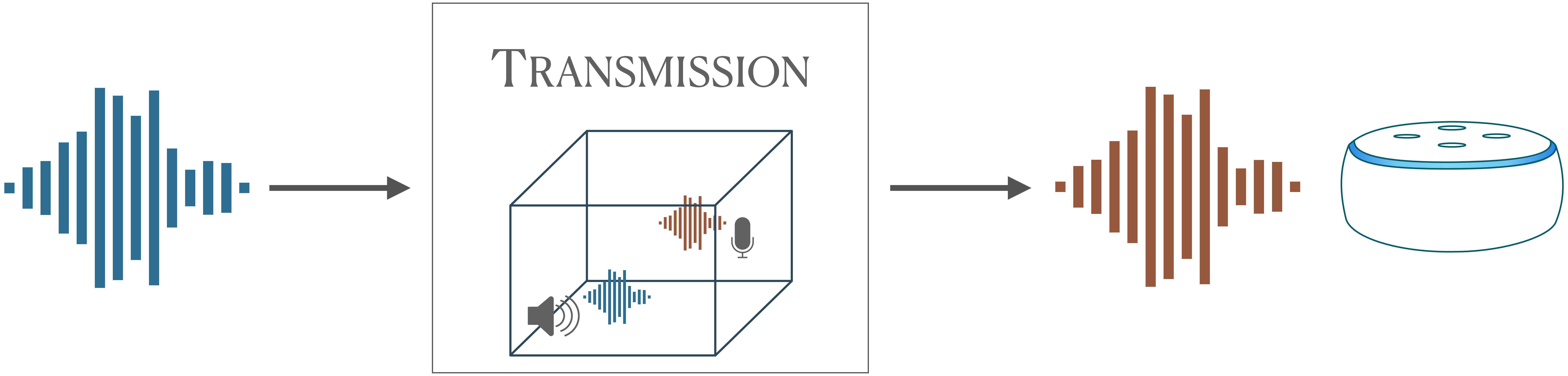


SPECIFICALLY THE UNION SAID IT WAS PROPOSING TO PURCHASE ALL OF THE ASSETS OF THE OF THE UNITED AIRLINES INCLUDING PLANES GATES FACILITIES AND LANDING RIGHTS

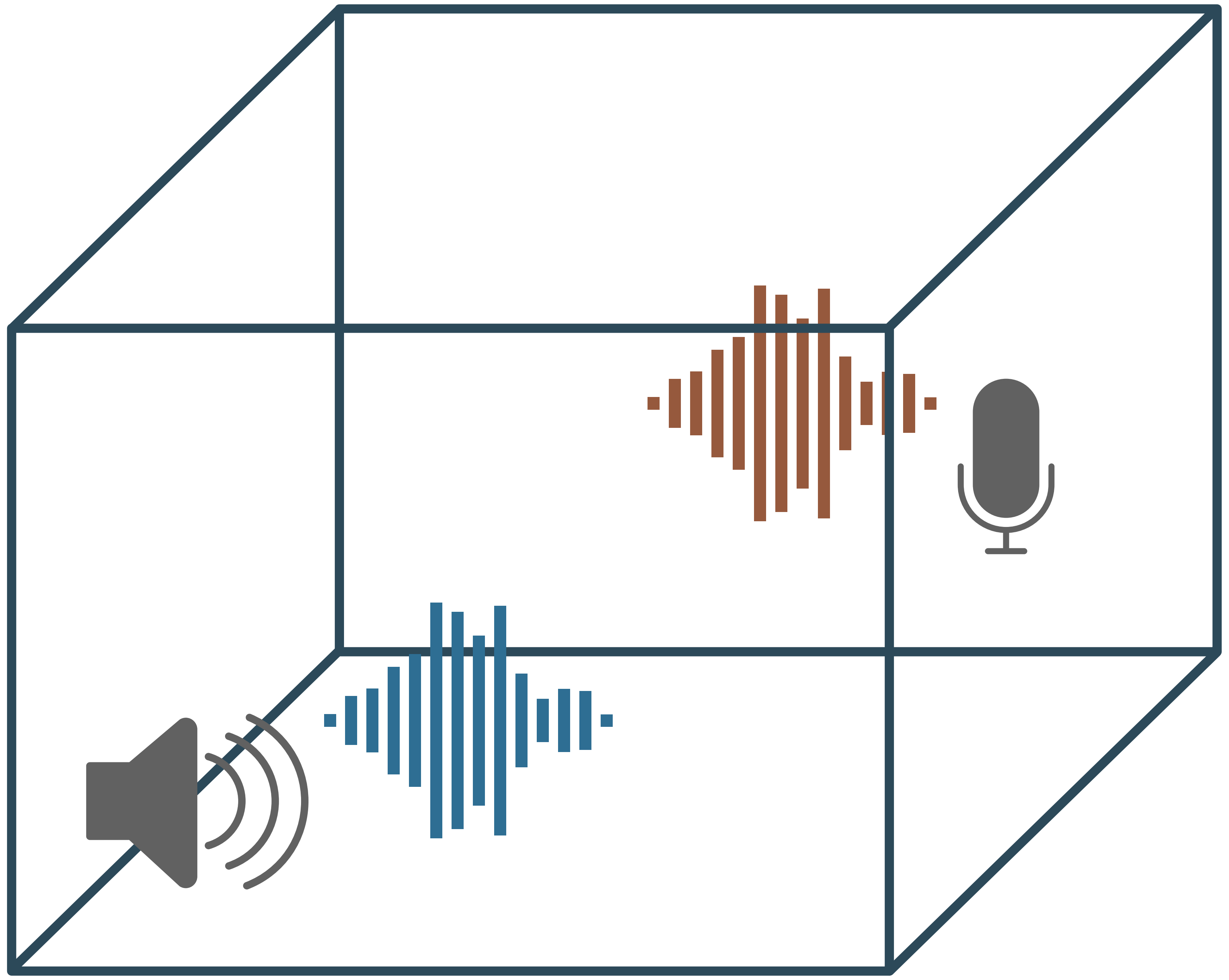
Adversarial



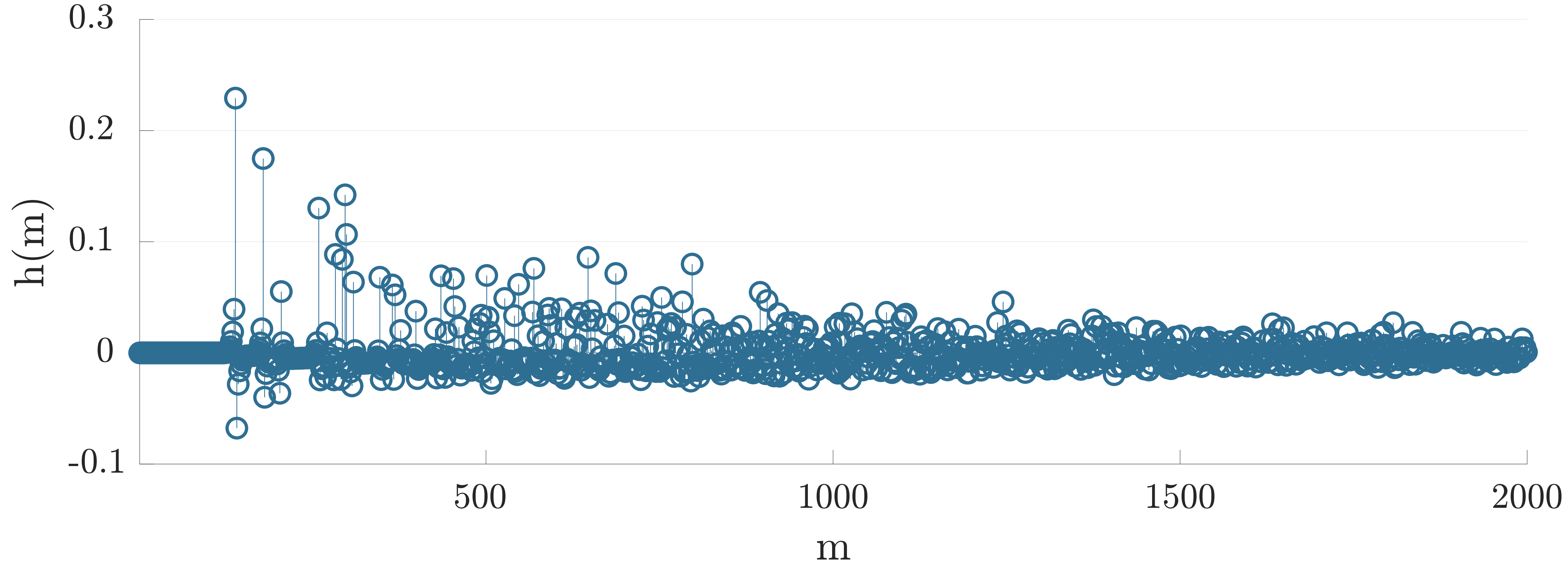
DEACTIVATE SECURITY CAMERA AND UNLOCK FRONT DOOR



Lea Schönherr, Thorsten Eisenhofer, Steffen Zeiler, Thorsten Holz, Dorothea Kolossa, “Imperio: Robust Over-the-Air Adversarial Examples for Automatic Speech Recognition Systems,” Annual Computer Security Applications Conference (ACSAC), 2020



ROOM IMPULSE RESPONSE (RIR)



$$x_h(n) = \sum_{m=0}^{M-1} x(n-m) \cdot h(m) \quad \forall\, n = 0,\ldots,N-1 \tag{7}$$

M: Length of the RIR

N: Length of the signal

Adversarial UNLOCK FRONT DOOR

Countermeasures

Audio Adversarial Examples

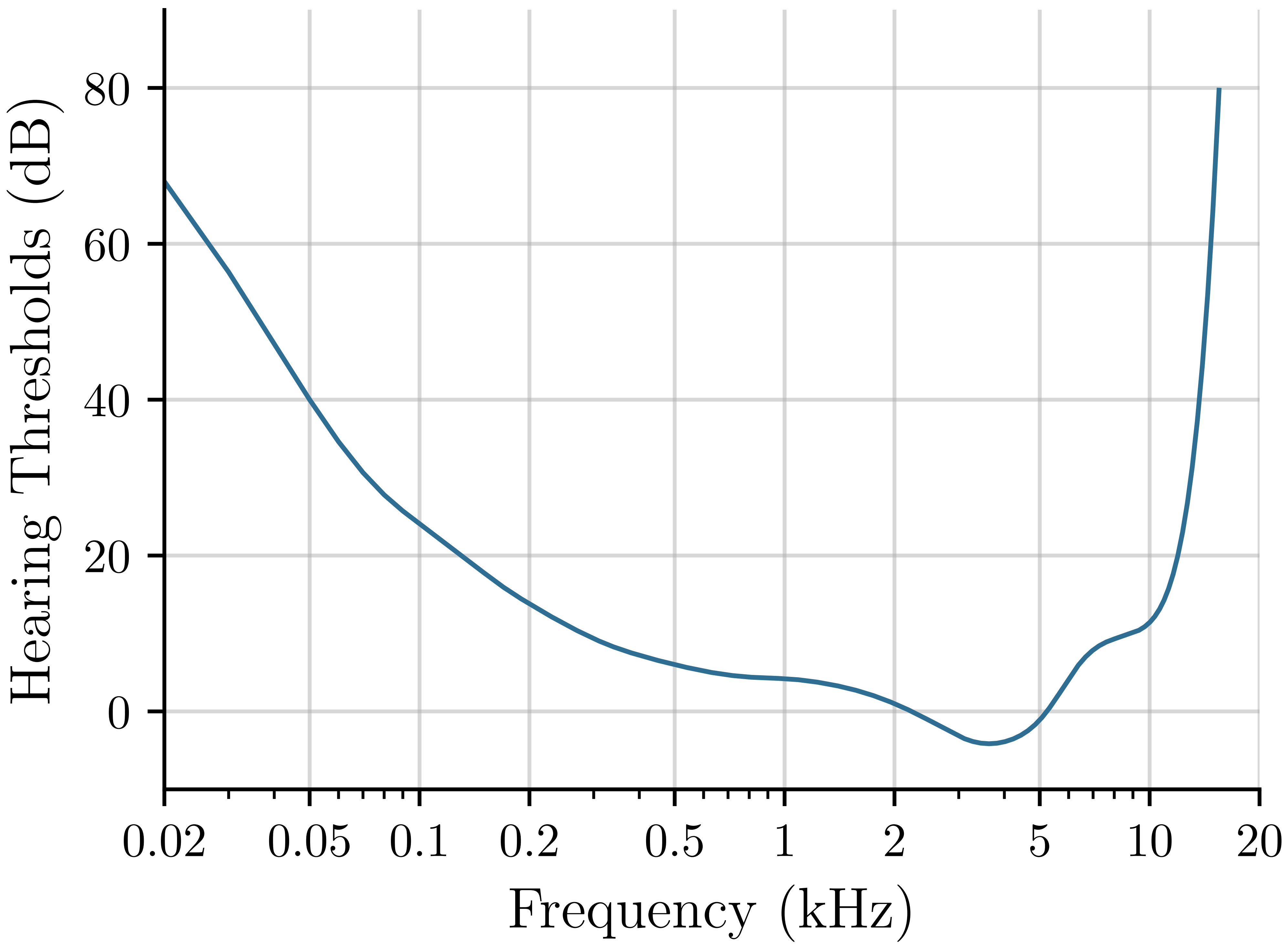
Thorsten Eisenhofer, Lea Schönherr, Joel Frank, Dorothea Kolossa, Thorsten Holz,
“Dompteur: Taming Audio Adversarial Examples”, USENIX Security Symposium, 2021.

Gustav Fechner

1801-1887

Gustav Fechner
1801-1887

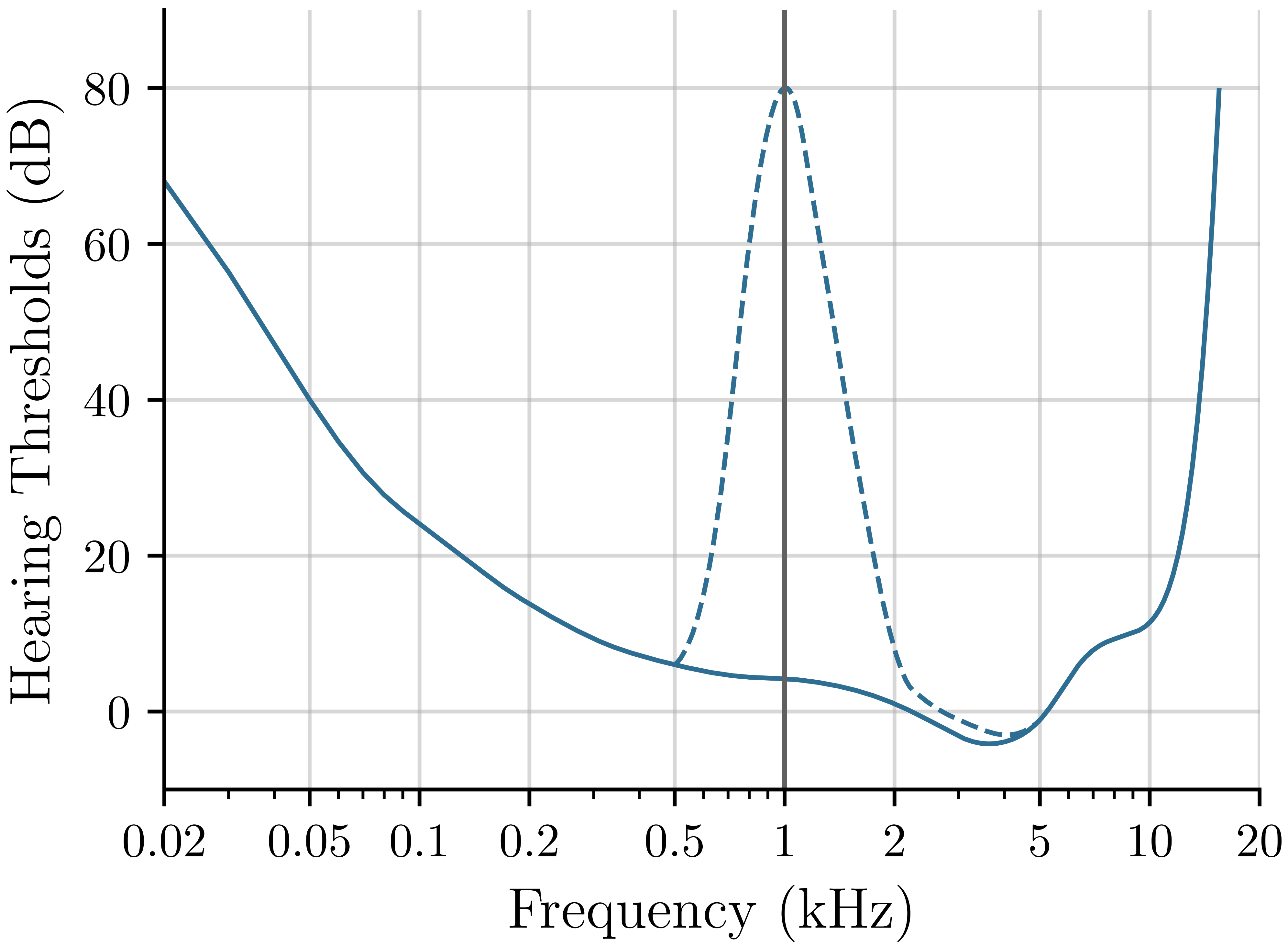
Absolute Hearing Thresholds



Gustav Fechner

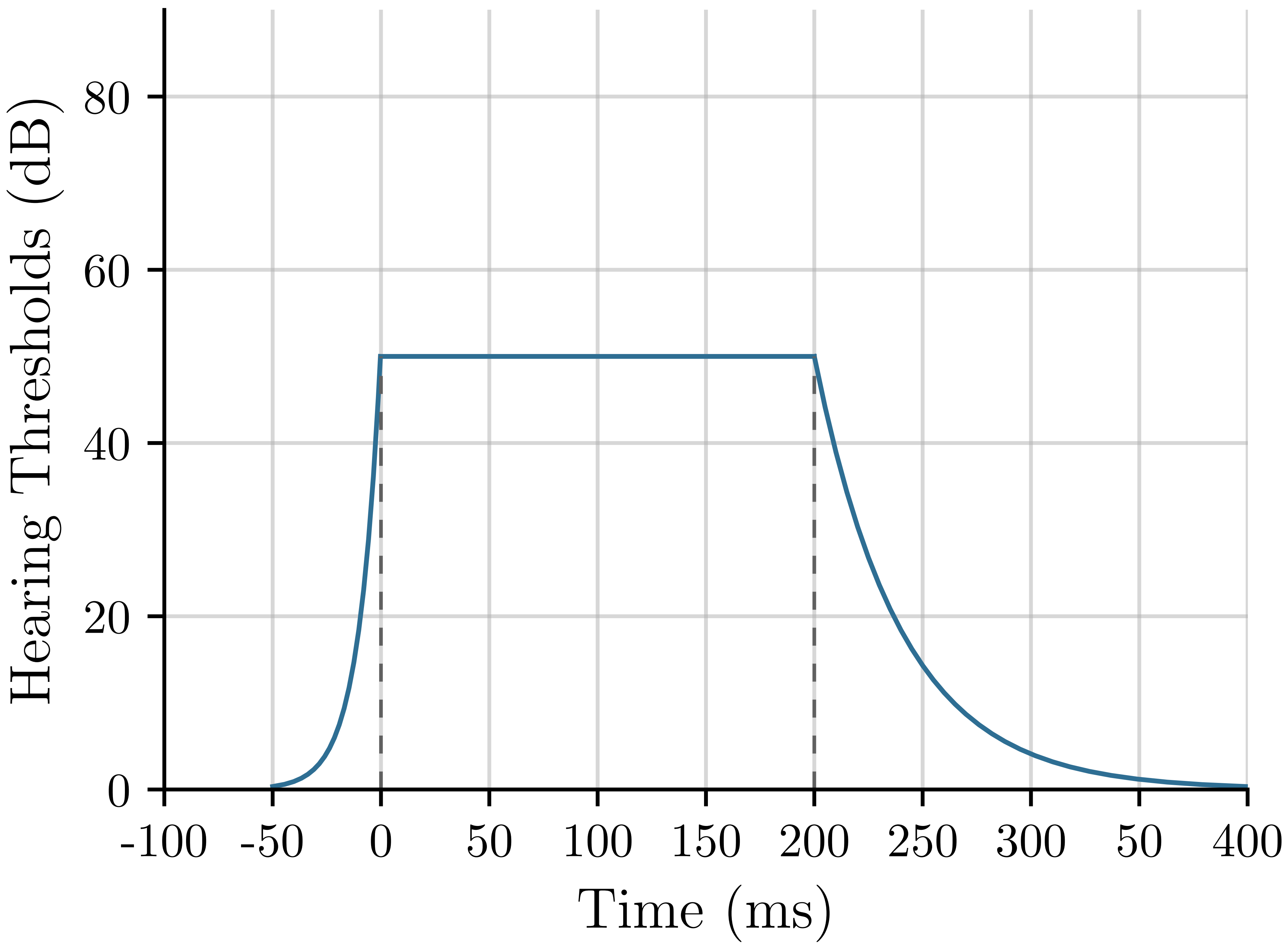
1801-1887

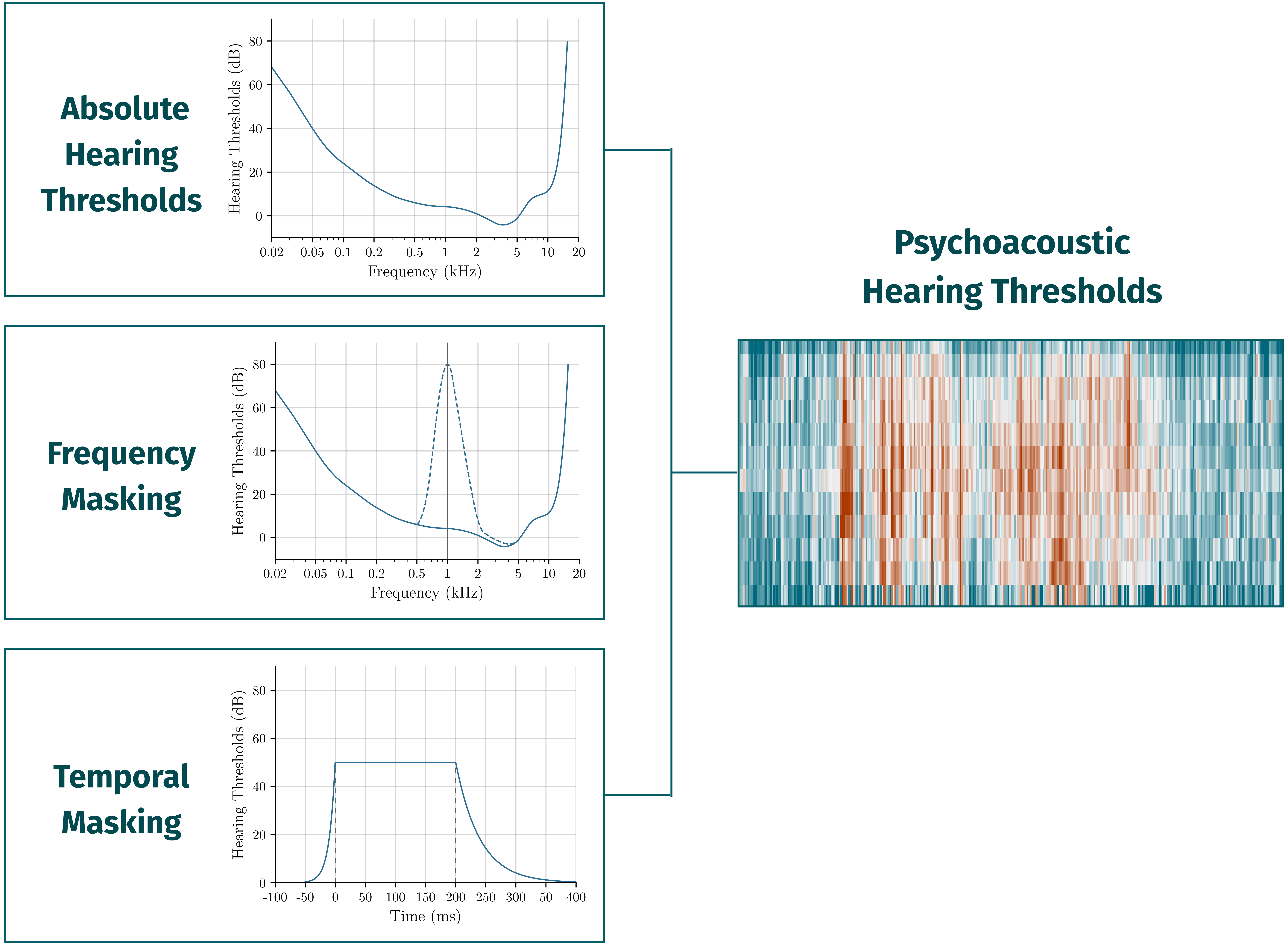
Frequency Masking



Gustav Fechner
1801-1887

Temporal Masking



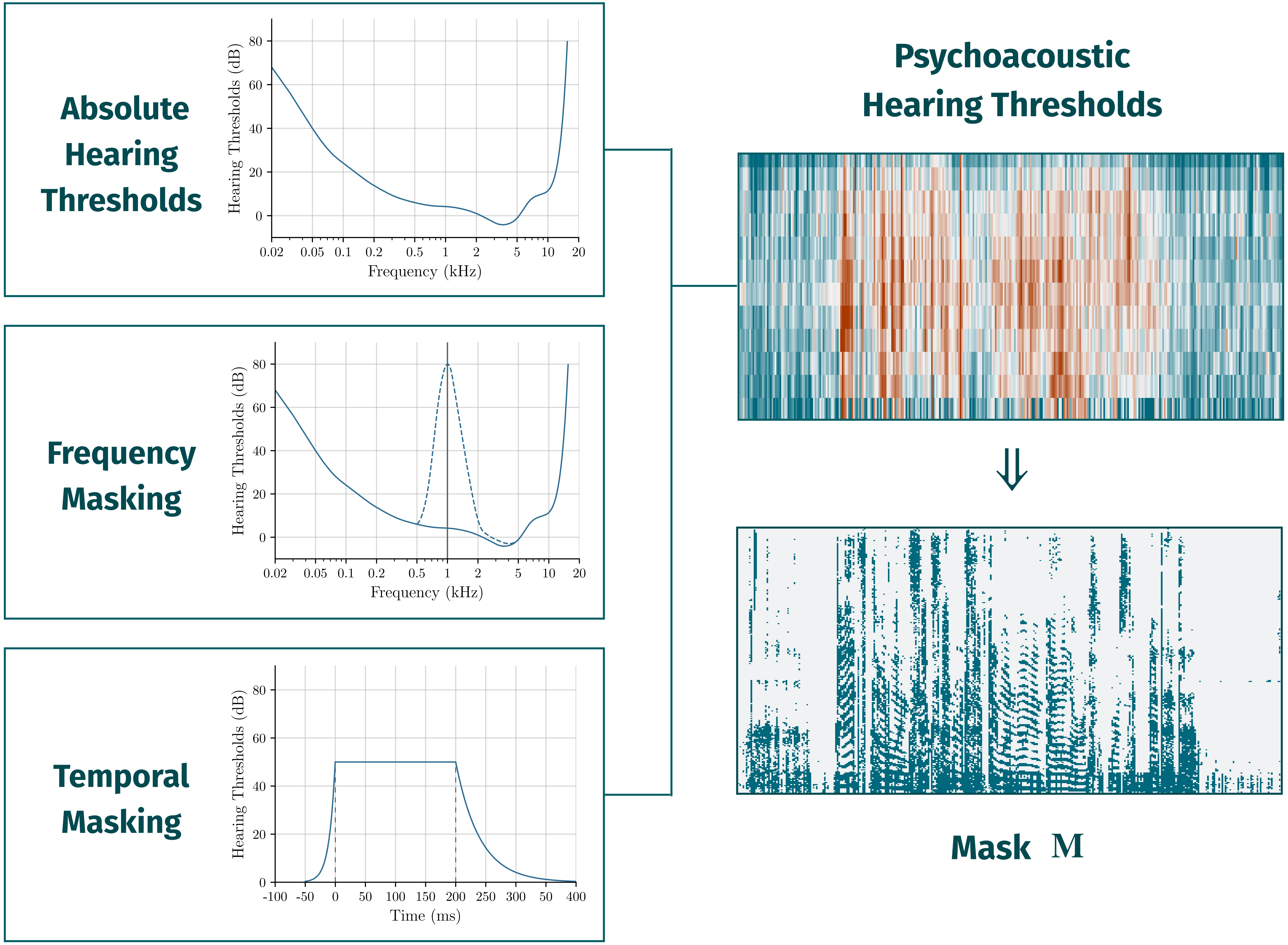


Absolute Hearing Thresholds

Frequency Masking

Temporal Masking

Psychoacoustic Hearing Thresholds



Absolute Hearing Thresholds

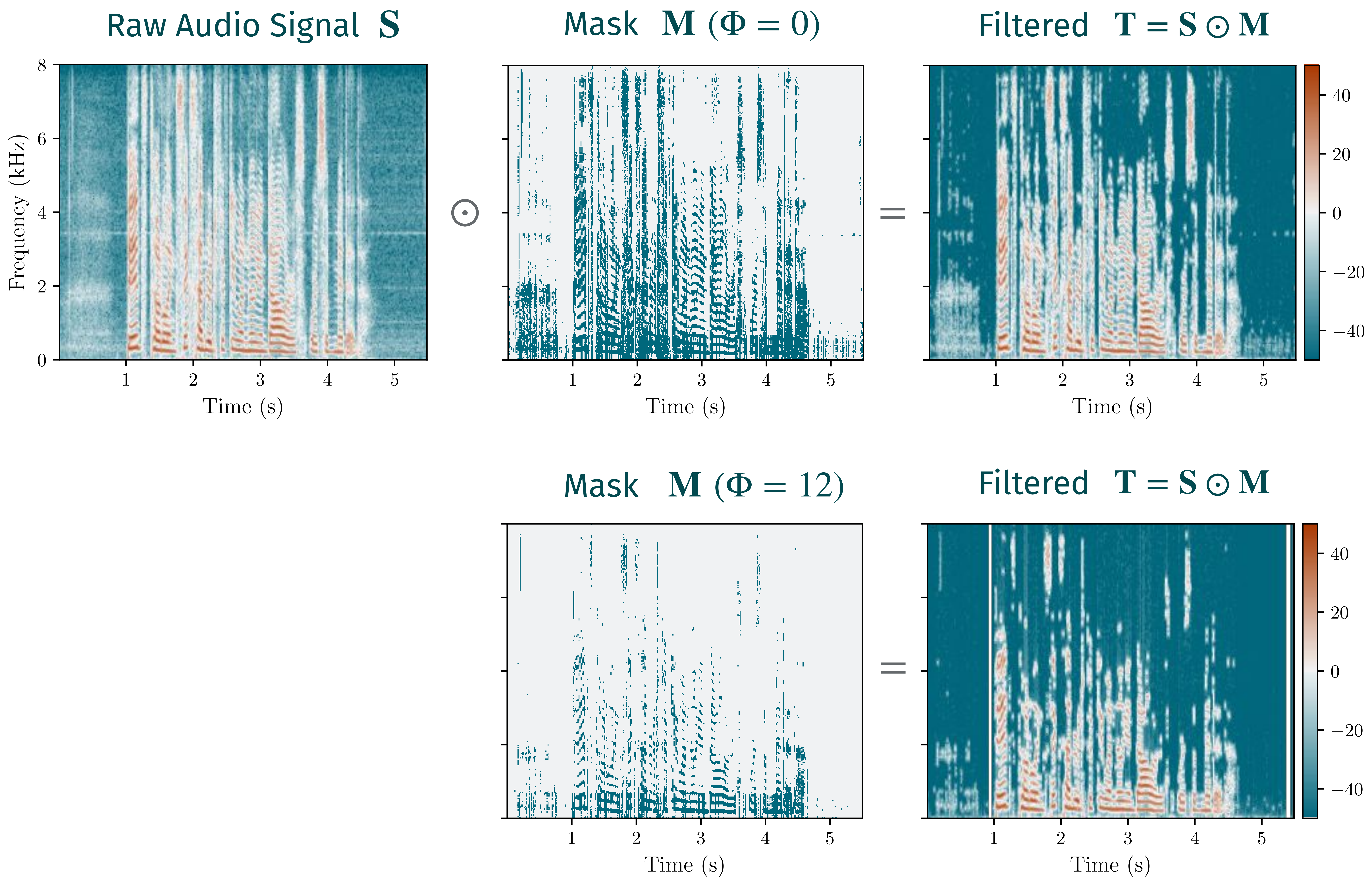
Frequency Masking

Temporal Masking

Psychoacoustic Hearing Thresholds

⇓

Mask $\mathbf{M}$



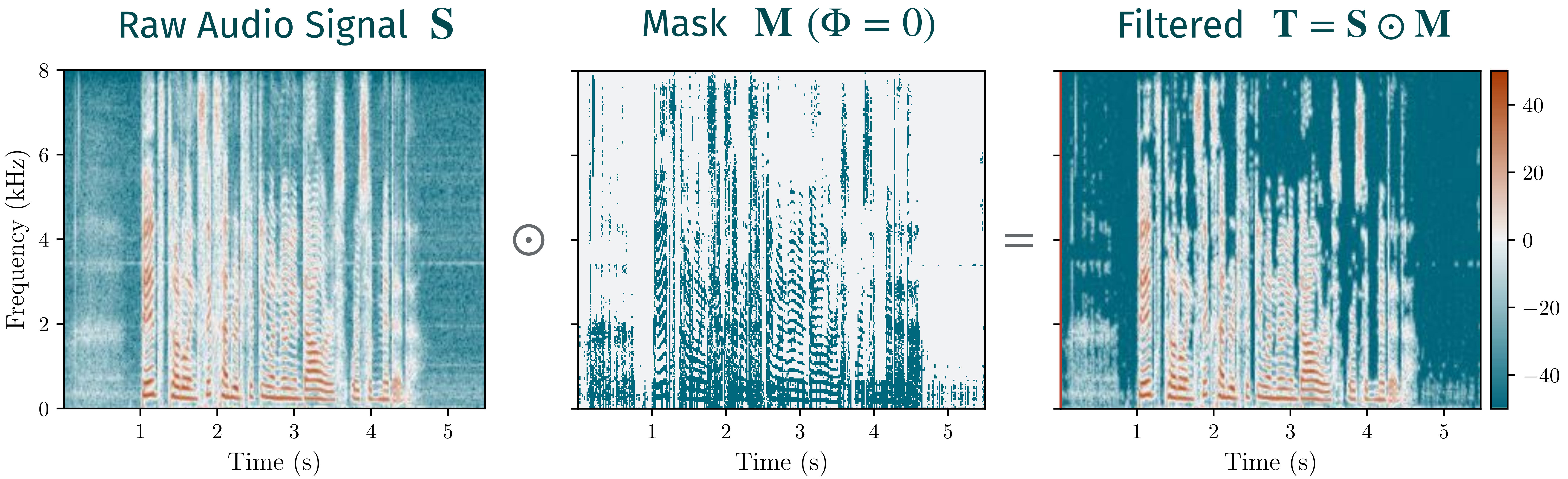
Raw Audio Signal $\mathbf{S}$

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

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

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

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

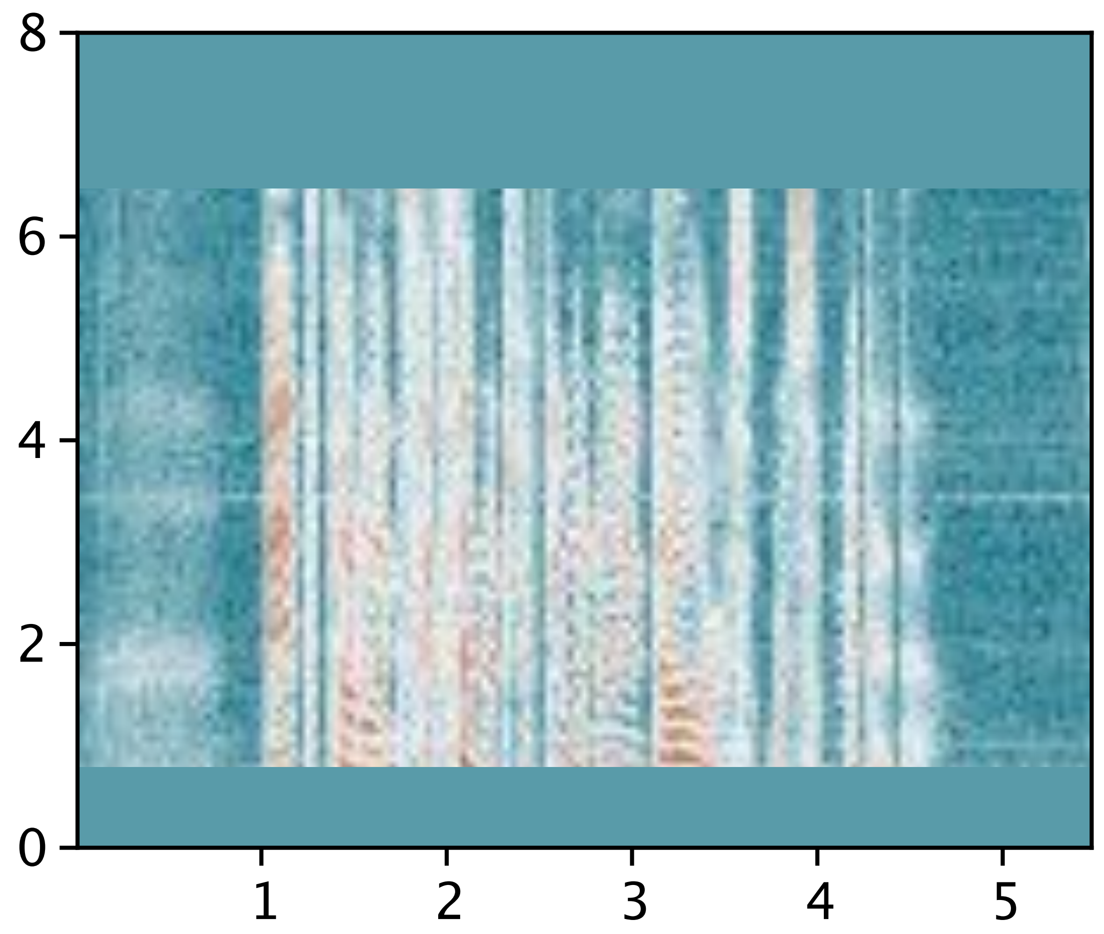


Raw Audio Signal $\mathbf{S}$

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

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

Band-Pass Filter

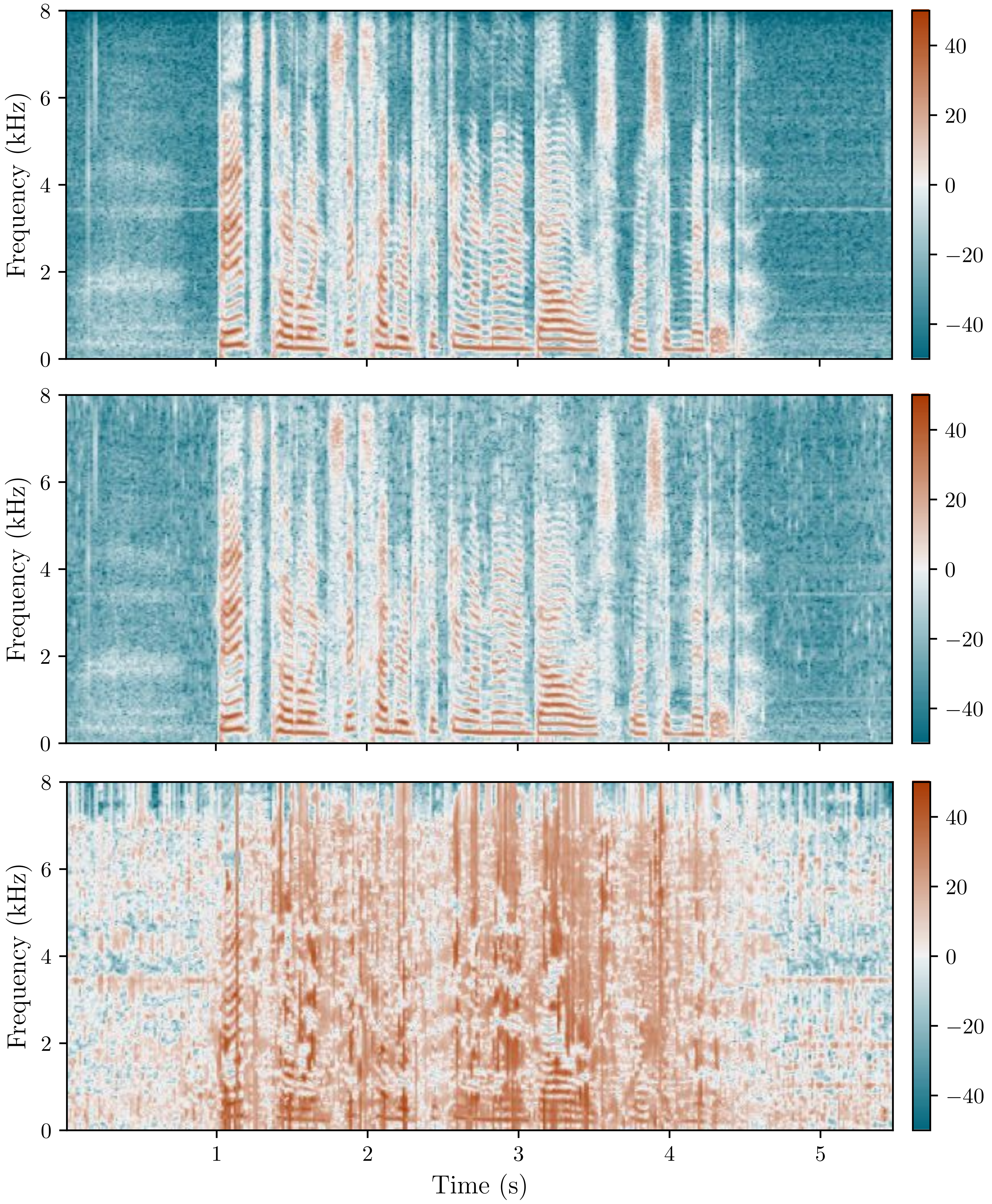


Adversarial Robustness

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

Previous Attack

SEND SECRET FINANCIAL REPORT

Dompteur $\Phi = 12$

SEND SECRET FINANCIAL REPORT

Large–Scale
Adversary

Robustness

- **Unacceptable**
  Computer Speech and Language (in submission)

Countermeasures

- **Dompteur**
  USENIX'21
- **Uncertainty**
  INTERSPEECH'20

Deep Fakes

- **WaveFake**
  NeurIPS'21 (in submission)
- **DeepFake**
  ICML'20

Adversarial Examples

- **Adversarial Attack**
  NDSS'19
- **Imperio**
  ACSAC'20

Data Poisoning

- **VenoMave**
  USENIX'22 (planned submission)

And more cool projects to come 🙌

LARGE-SCALE
ADVERSARY

Speech recognition not robust

Attacks possible both during runtime and training time

Psychoacoustics effective to force attacker into audible ranges

Thank you!