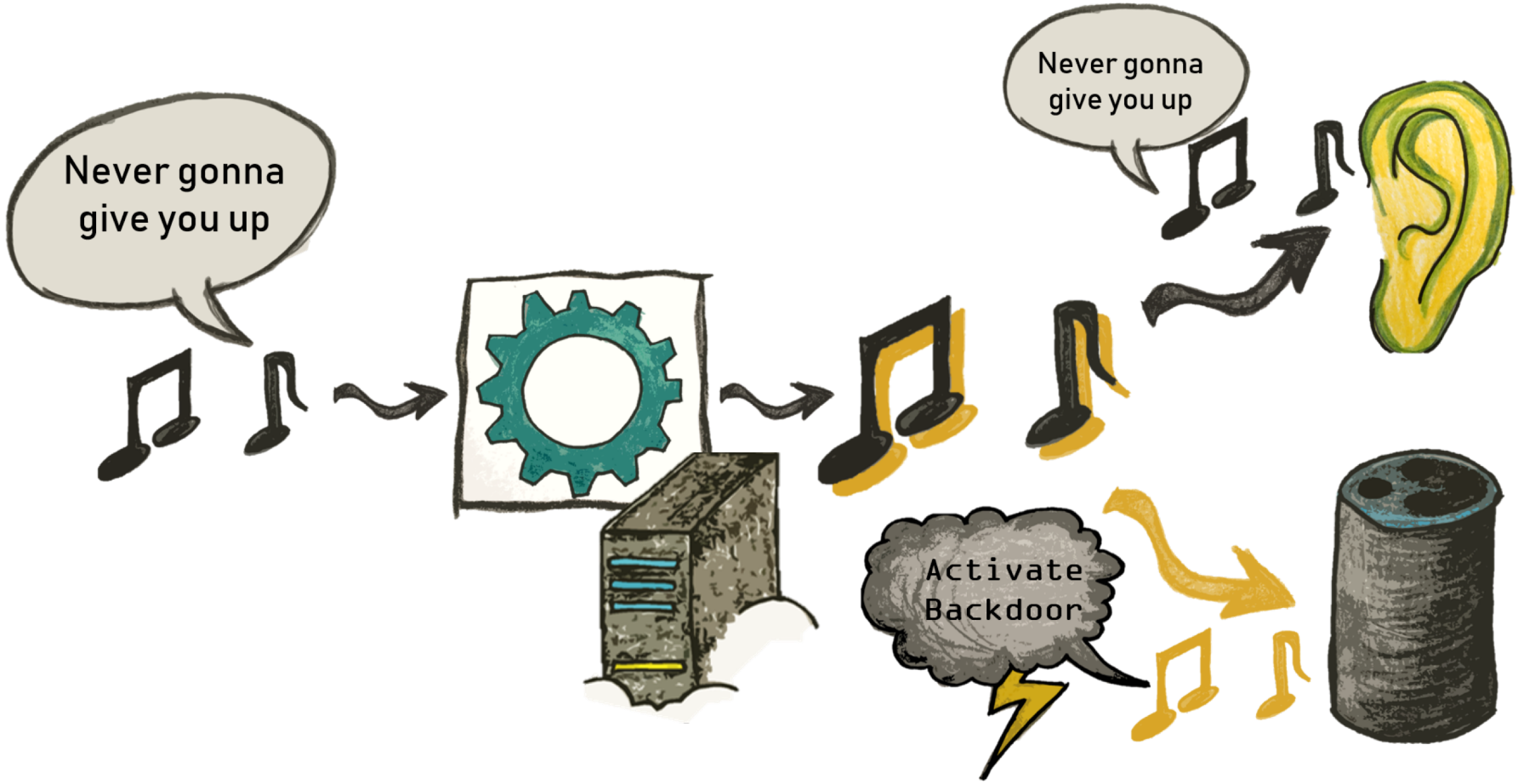
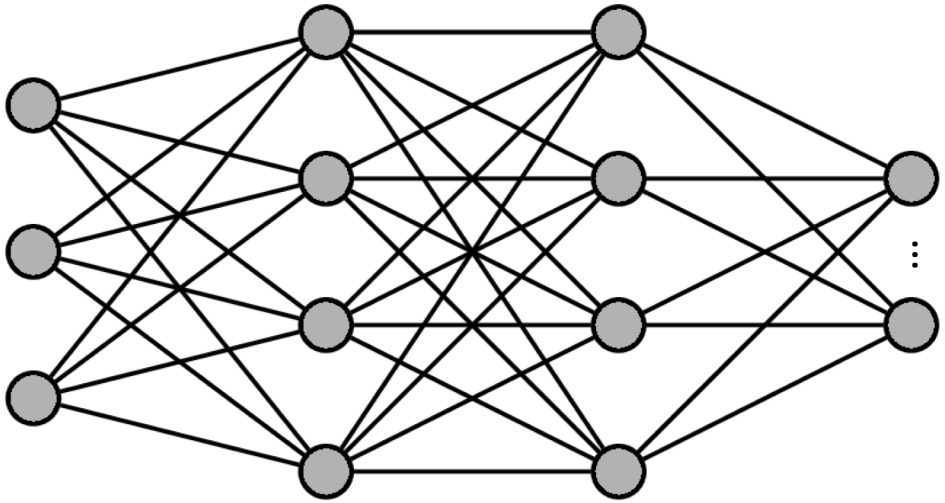


Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding

Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, Dorothea Kolossa – Ruhr University Bochum

Adversarial Machine Learning

Input x:



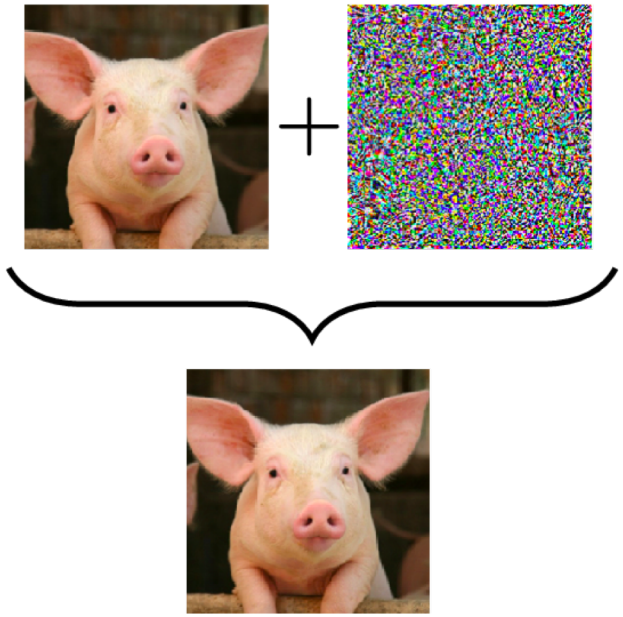
$$y = F(x)$$

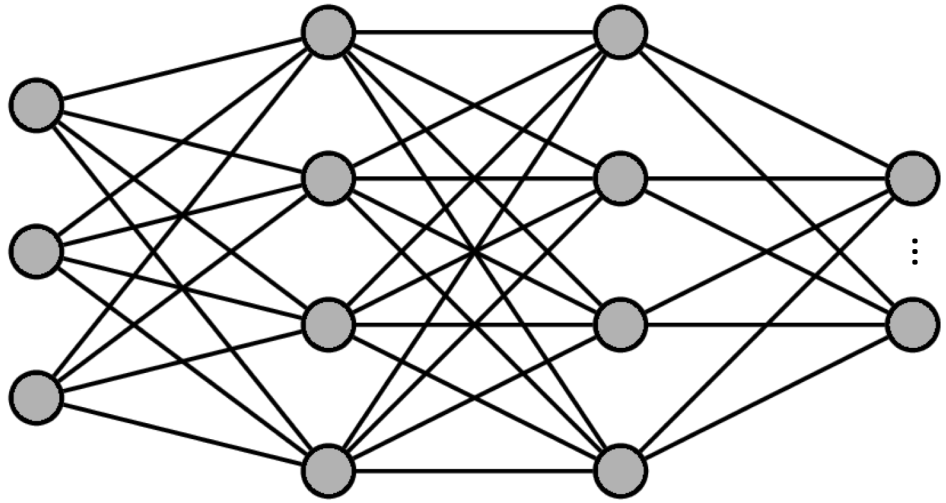
Output y:

Pig

Adversarial Machine Learning

Input $x + \delta$:





$$y = F(x + \delta)$$

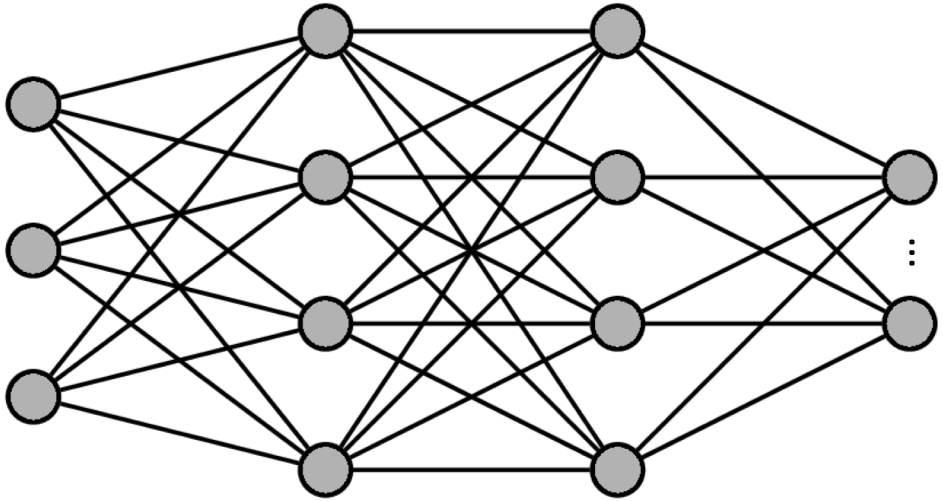
Output y:

Airliner

Adversarial Machine Learning

Input $x + \delta$:

$+ \delta$



$$y = F(x + \delta)$$

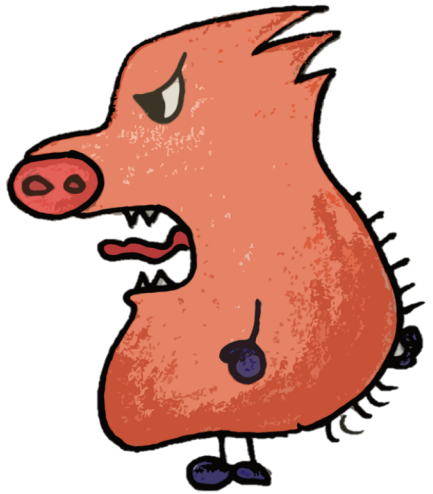
Output y:

DEACTIVATE SECURITY CAMERA AND UNLOCK FRONT DOOR

- For automatic speech recognition, the audio signal will be transcribed into the target text

Threat Model

- We assume a white-box attack
- The speech recognition system is trained to give the best possible recognition rate
- We assume a perfect transmission channel
- We only consider targeted attacks

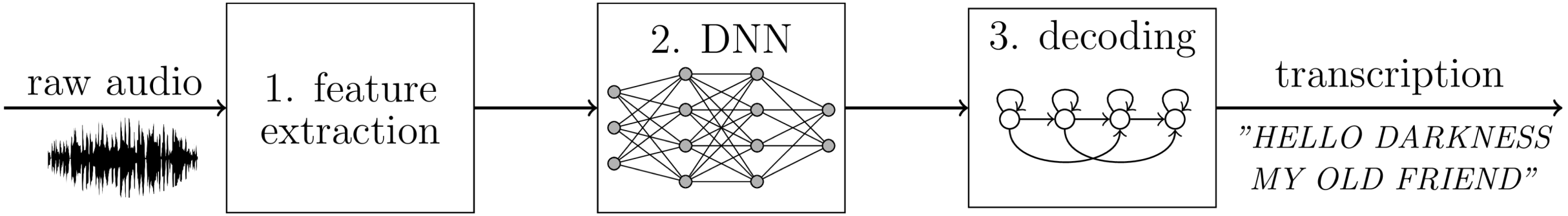


Speech Recognition System

Feature Extraction, DNN, and Decoding

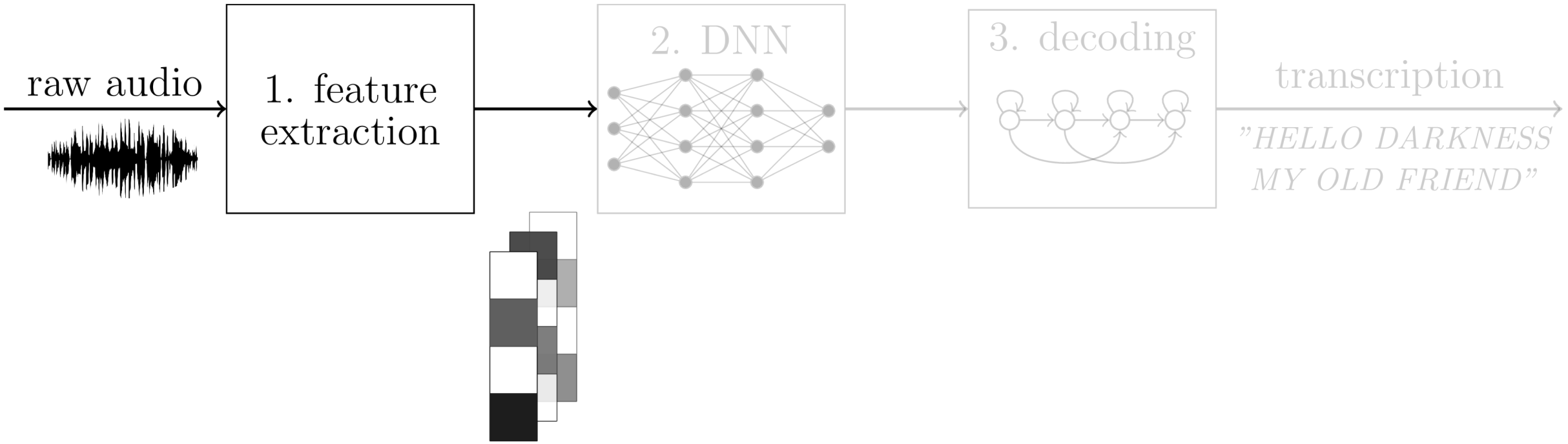
DNN-HMM Hybrid Automatic Speech Recognition

Based on the state-of-the-art Kaldi[1] toolkit



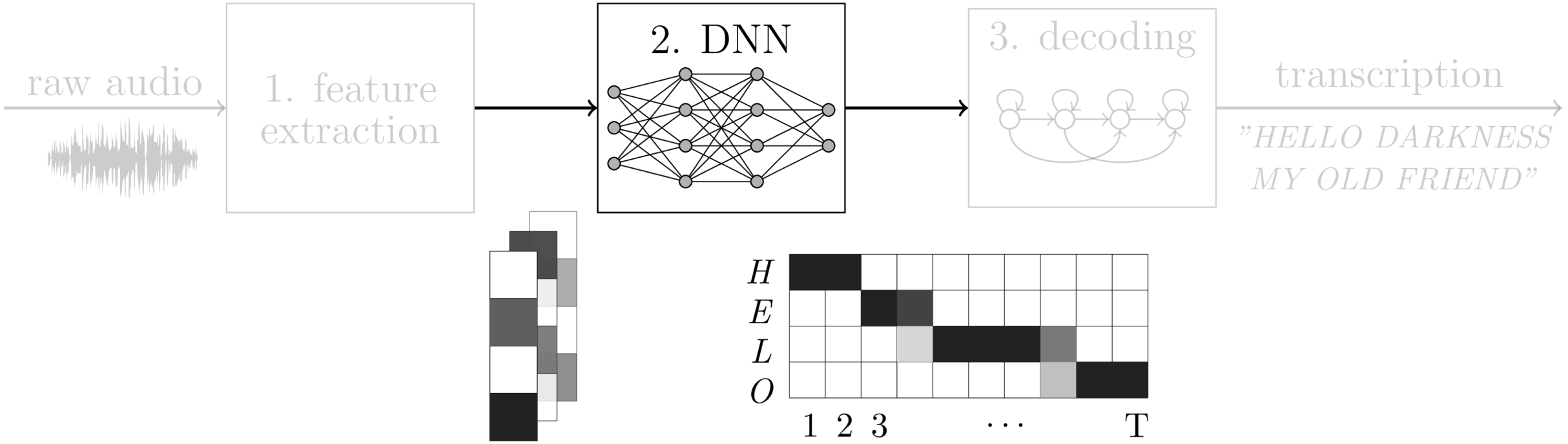
[1] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.

Automatic Speech Recognition – Feature Extraction



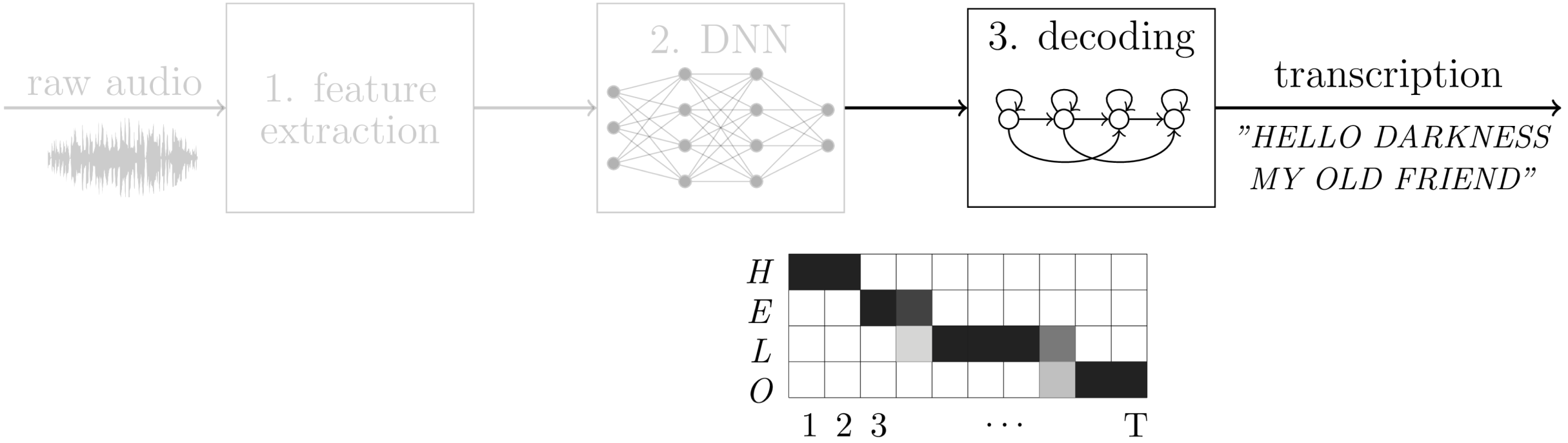
- The feature extraction calculates features in the frequency domain.

Automatic Speech Recognition - DNN



- The DNN maps the features to a matrix, describing the probability for each phone in each time step.

Automatic Speech Recognition - Decoding

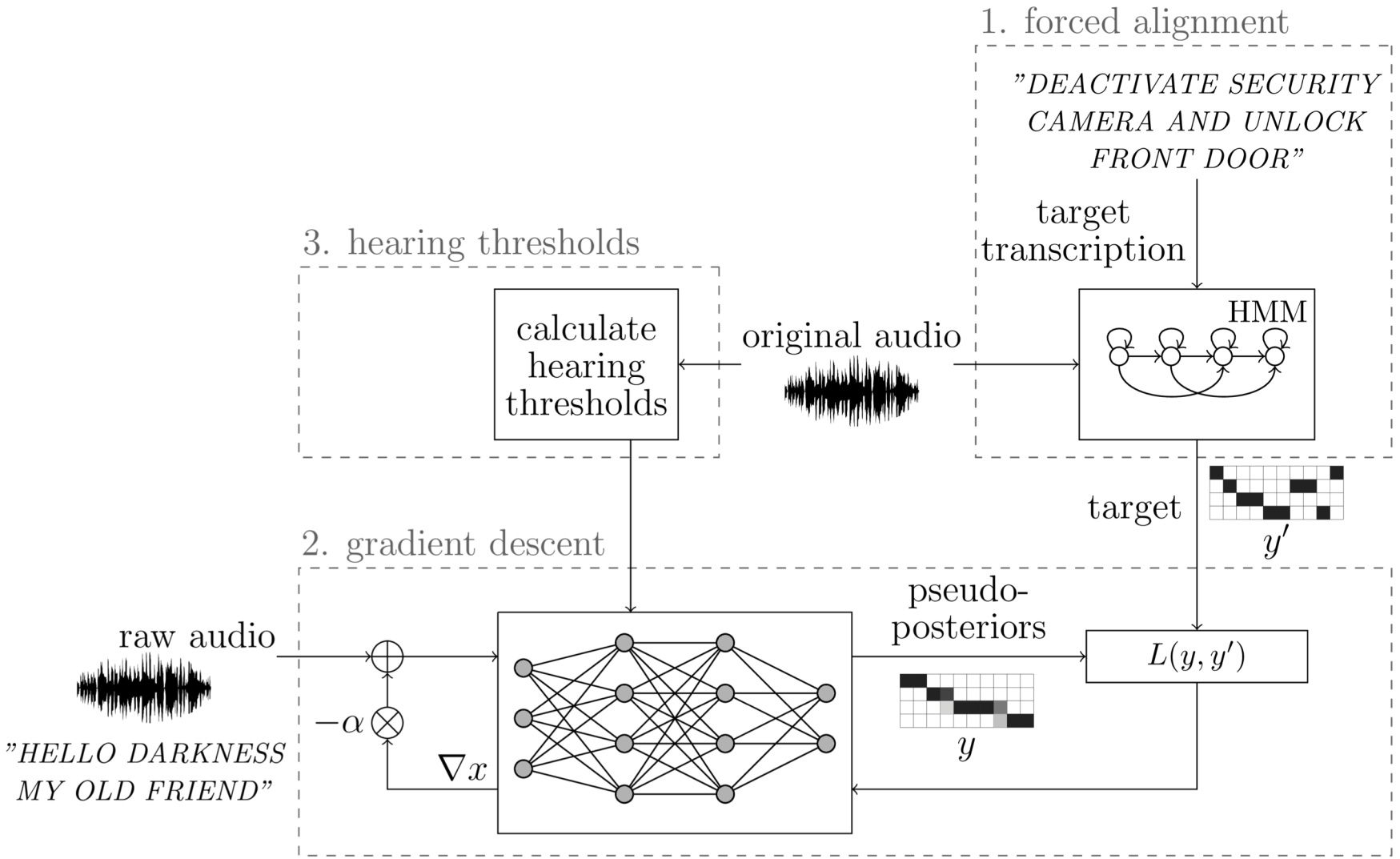


- The output of the DNN is used to find the most likely transcription with the underlying hidden Markov Model (HMM).

Attacking Speech Recognition

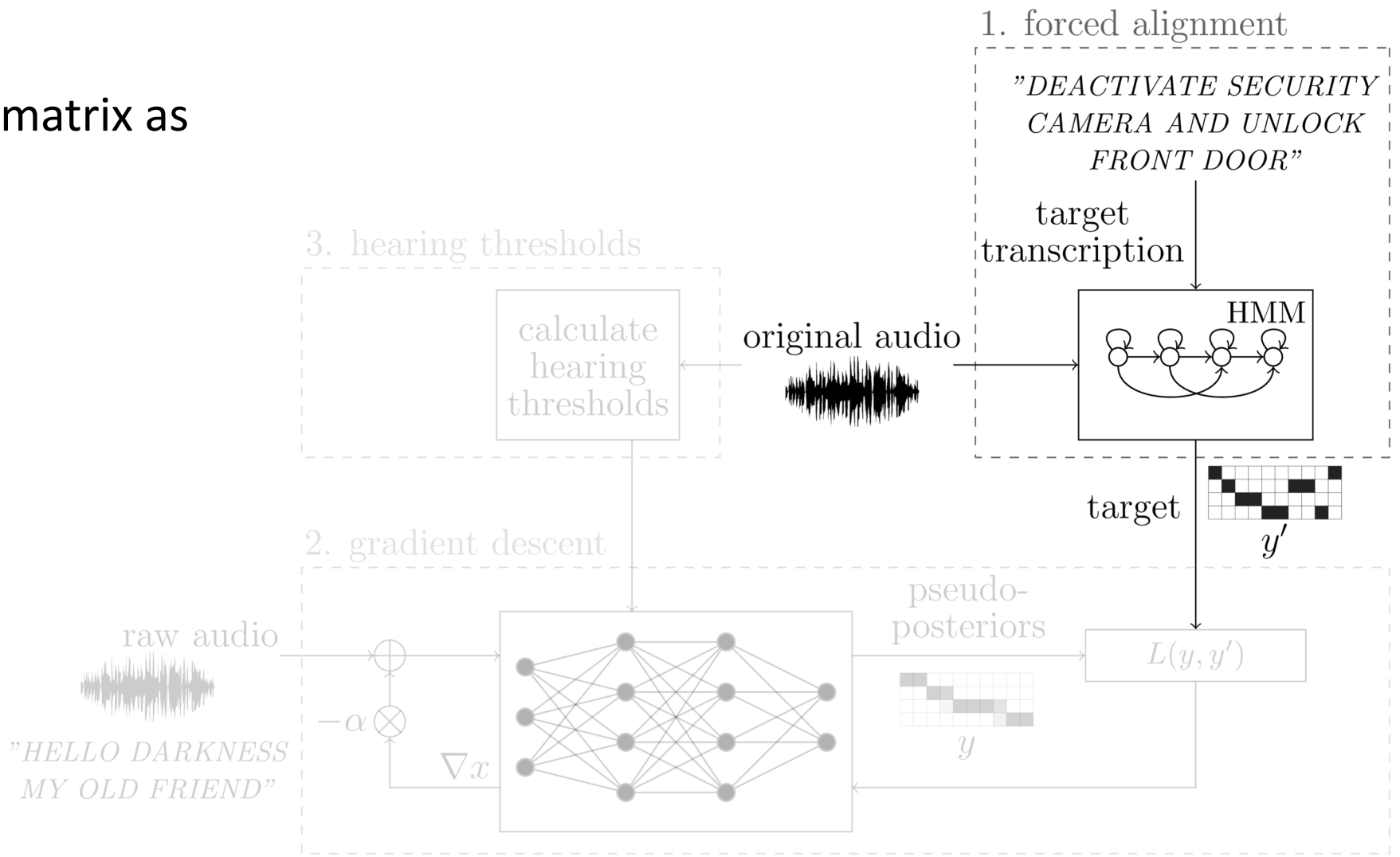
Forced Alignment, Gradient Descent, and Psychoacoustics

Attacking Speech Recognition



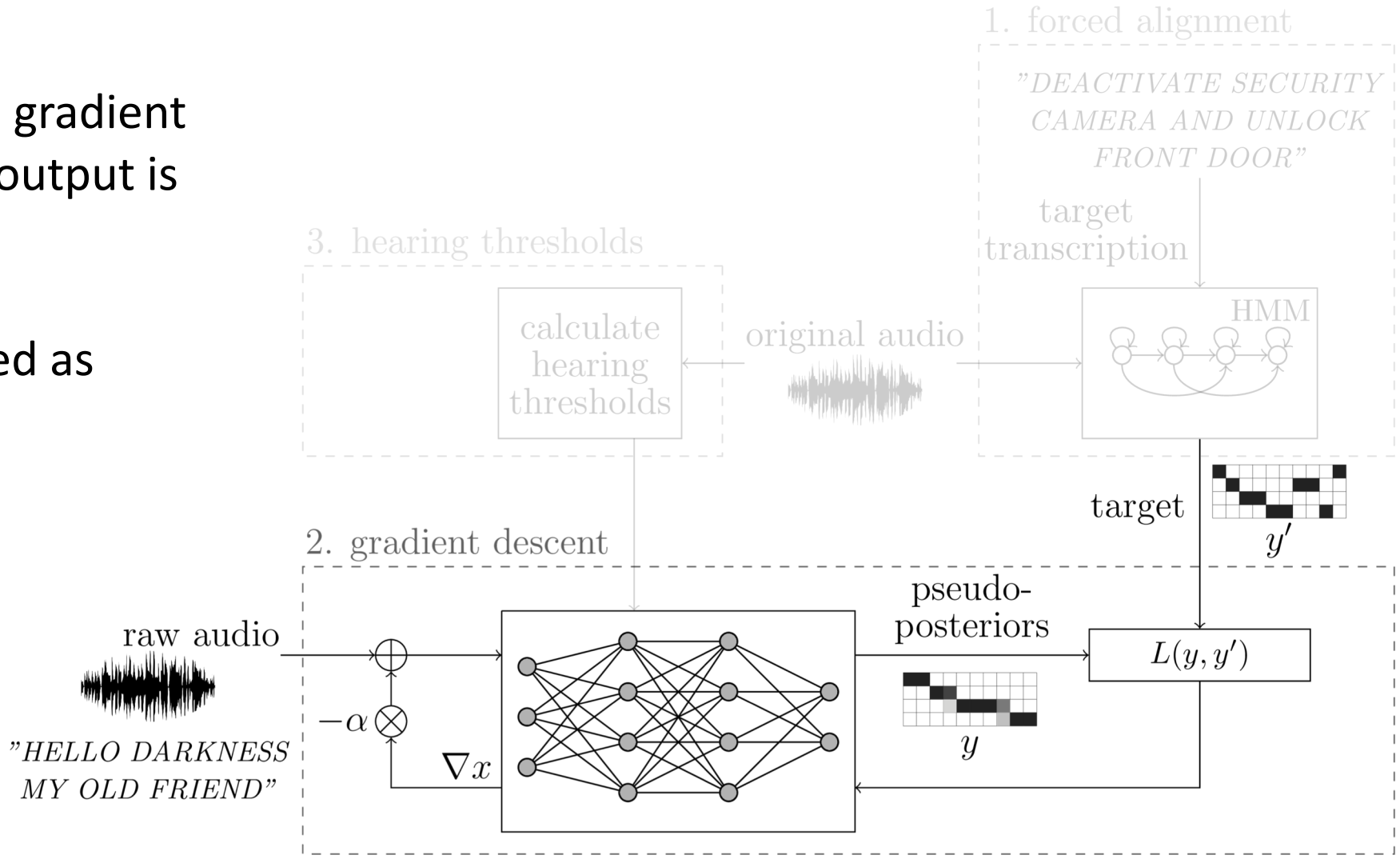
1. Forced Alignment

- Finds the best posterior matrix as the target for the DNN

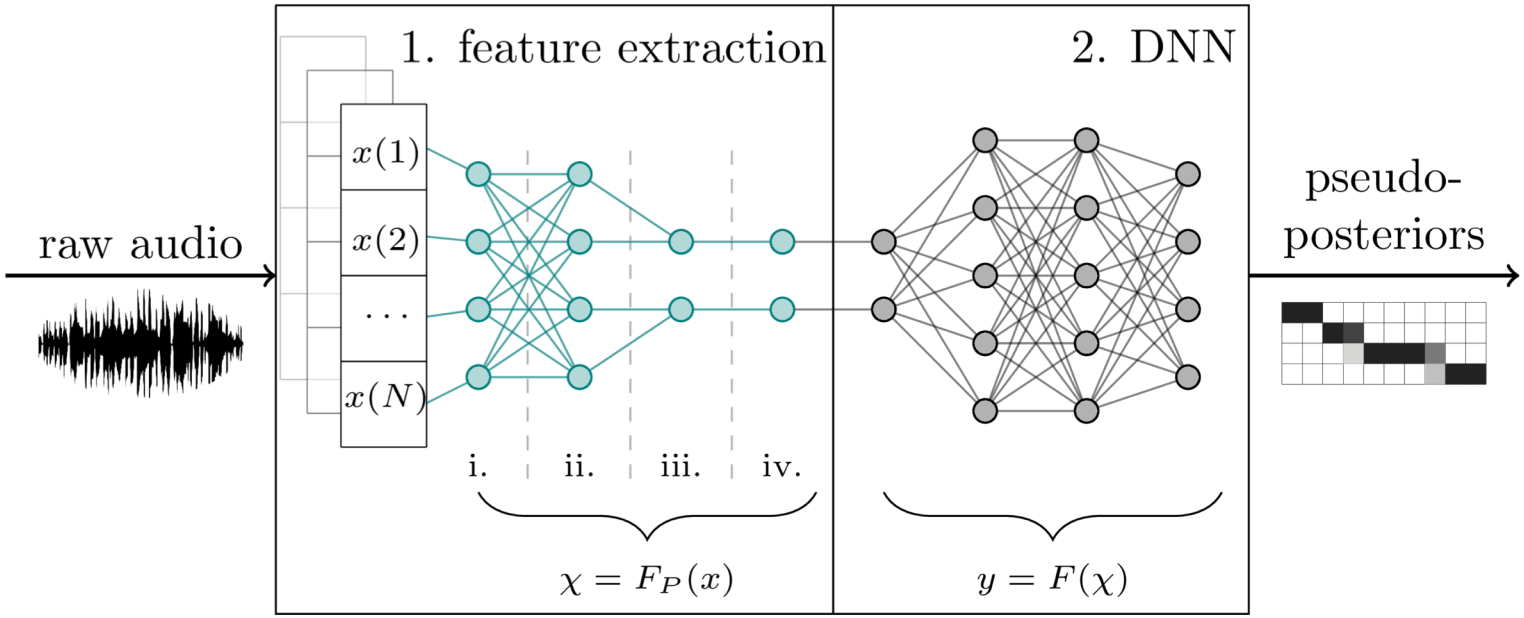


2. Gradient Descent

- The audio is updated via gradient descent until the target output is obtained
- The loss $L(y, y')$ is defined as cross-entropy



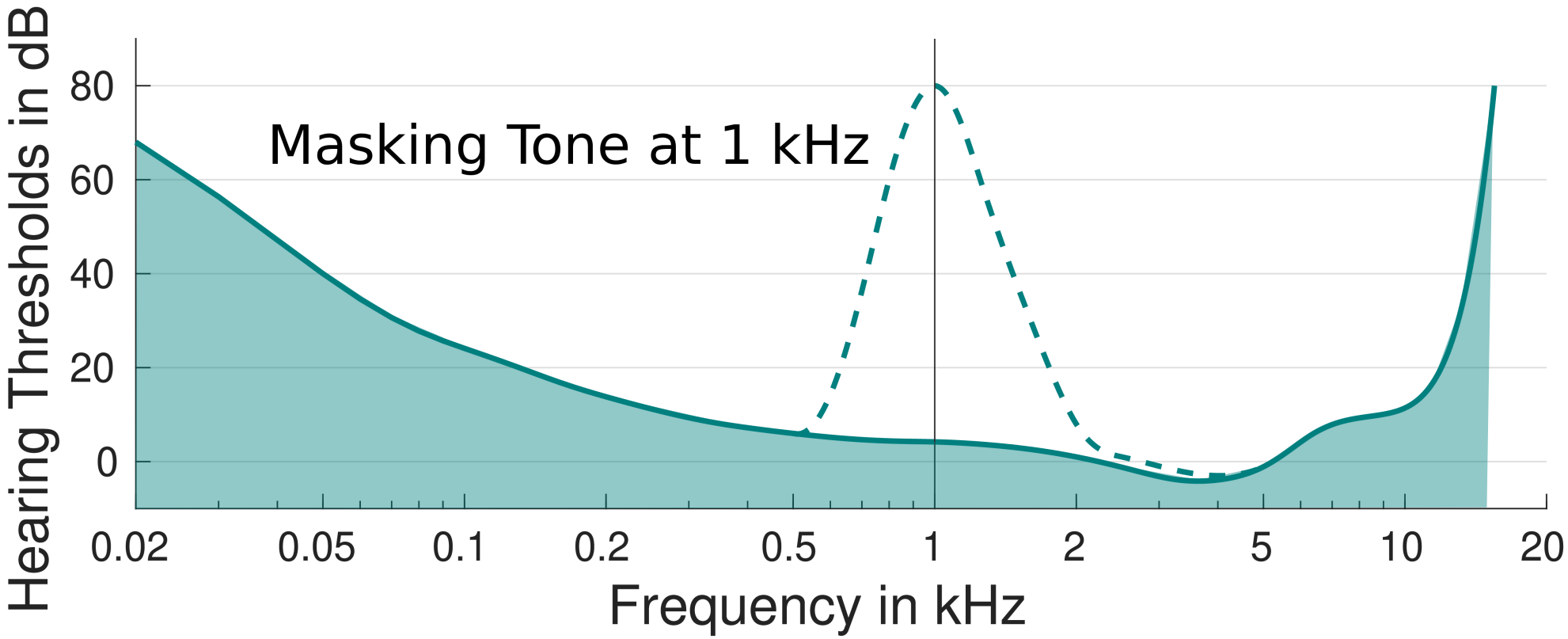
Integration of the Feature Extraction



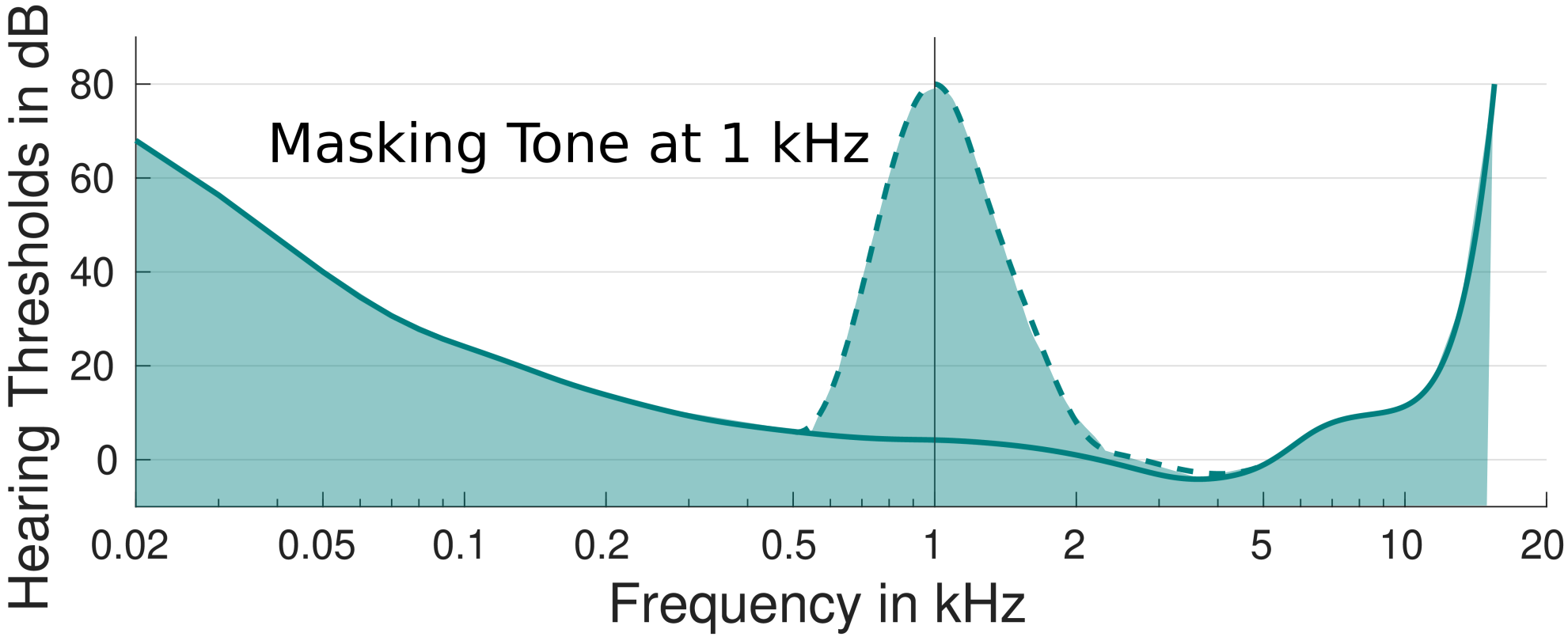
- The feature extraction $\chi = F_P(x)$ is integrated into the DNN $y = F(\chi)$
- This allows to update the raw audio directly

i. Framing and Window Function
ii. Discrete Fourier Transform
iii. Magnitude
iv. Logarithm

Psychoacoustics – Frequency Masking

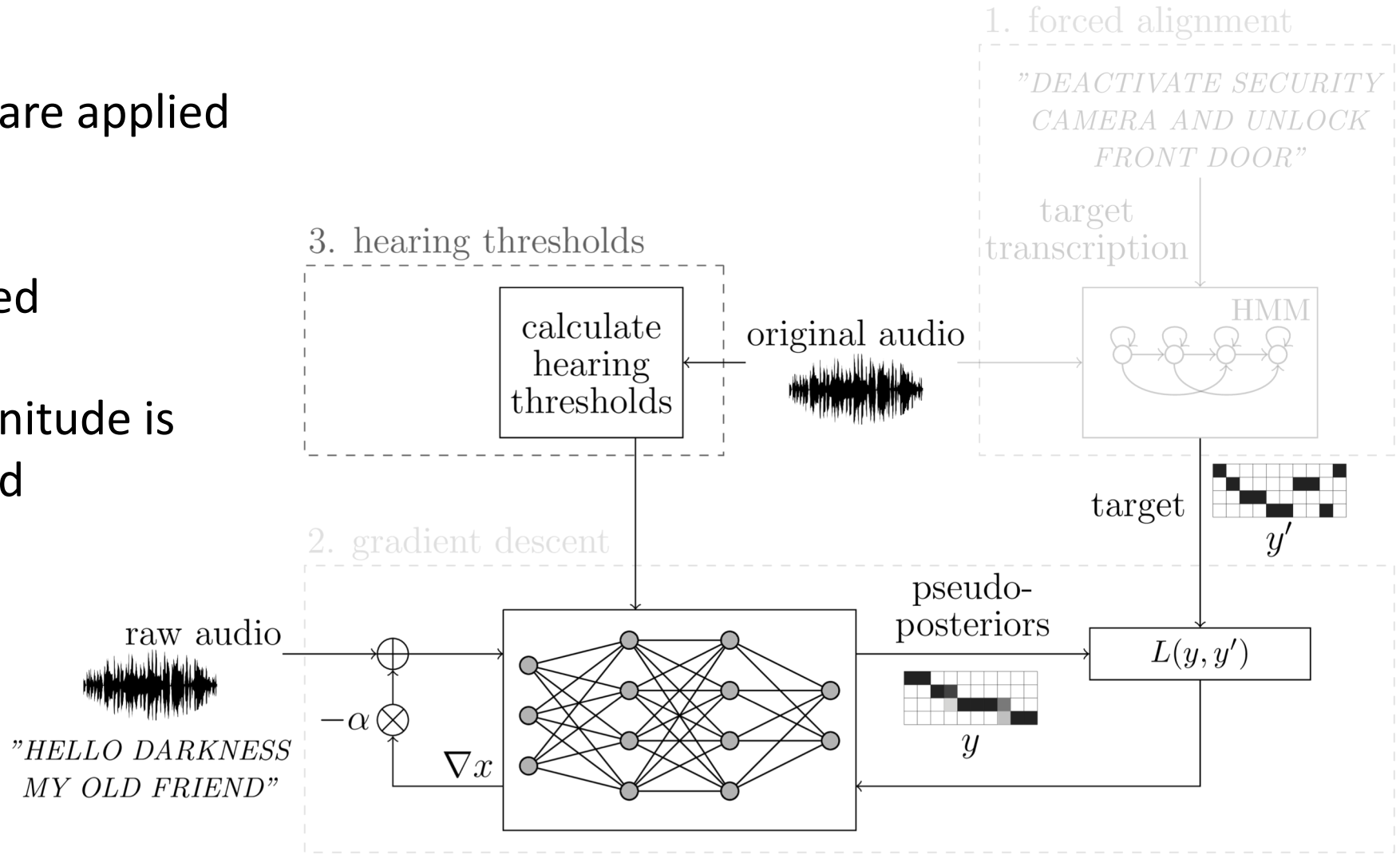


Psychoacoustics – Frequency Masking



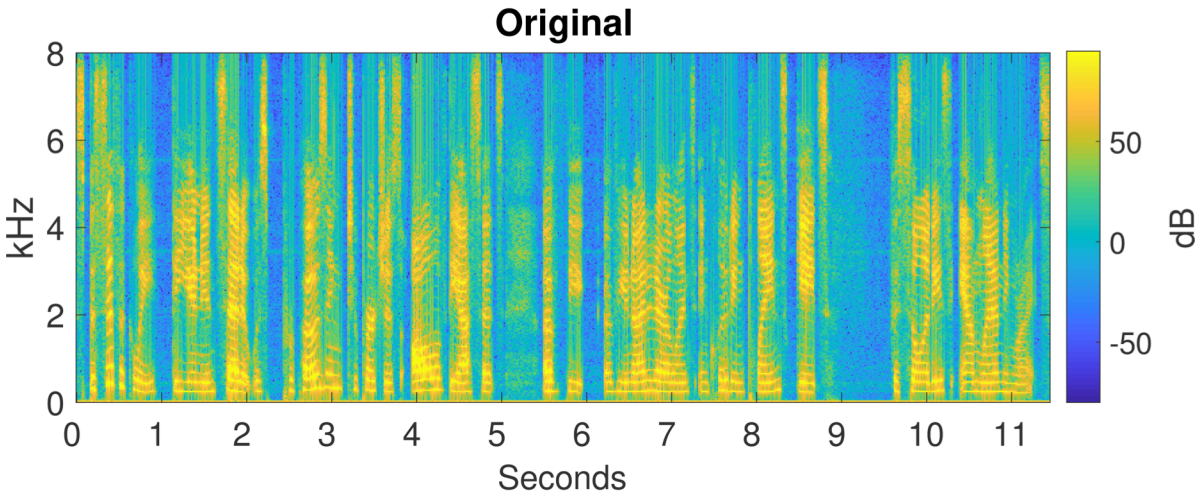
3. Hearing Thresholds

- The hearing thresholds are applied to limit the changes
- The MP3 principle is used
- The gradient of the magnitude is scaled with the threshold

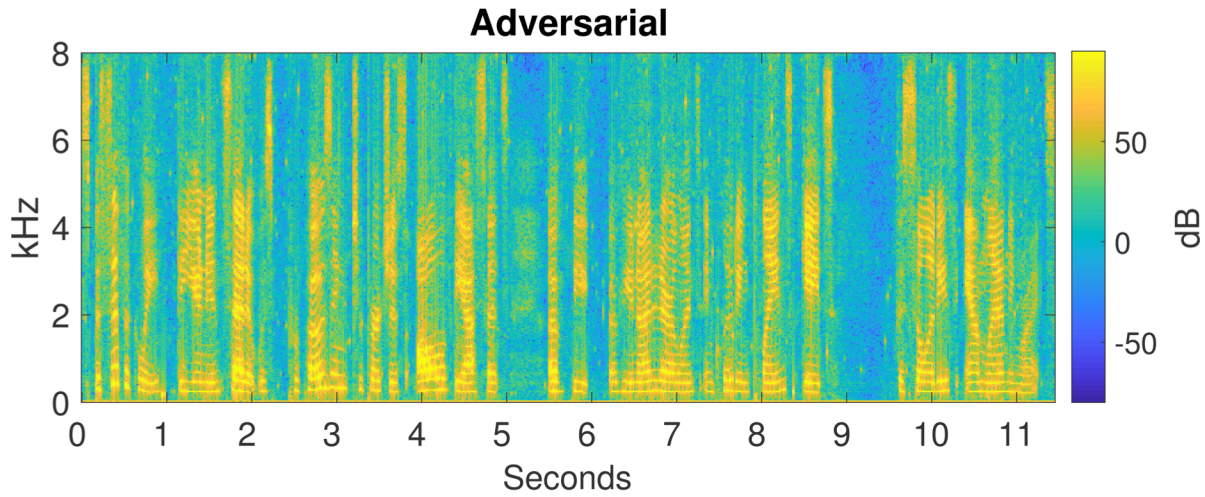


Results

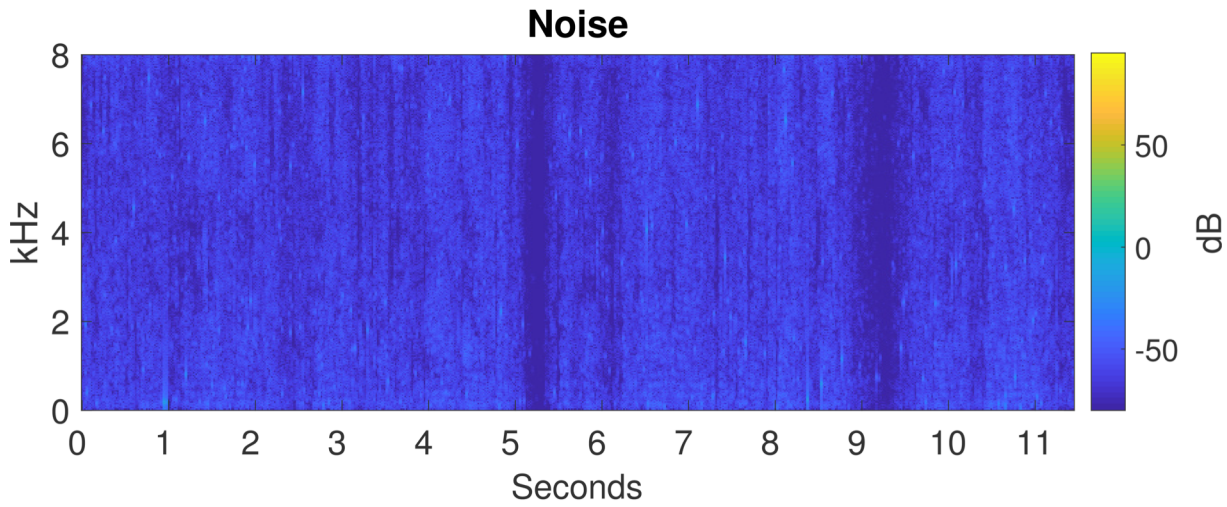
Audio Examples, Performance Analysis, and Listening Test



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



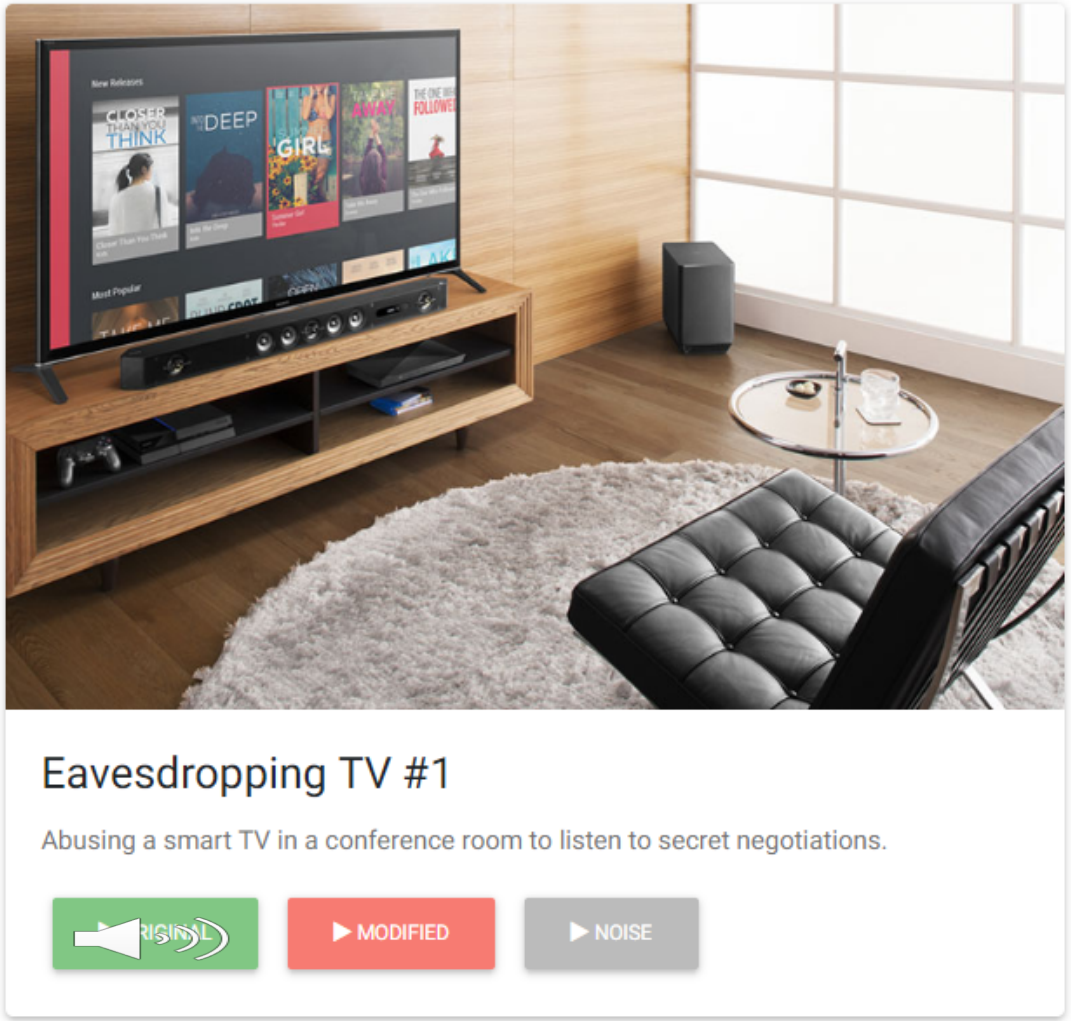
DEACTIVATE SECURITY CAMERA AND UNLOCK FRONT DOOR



Example 1

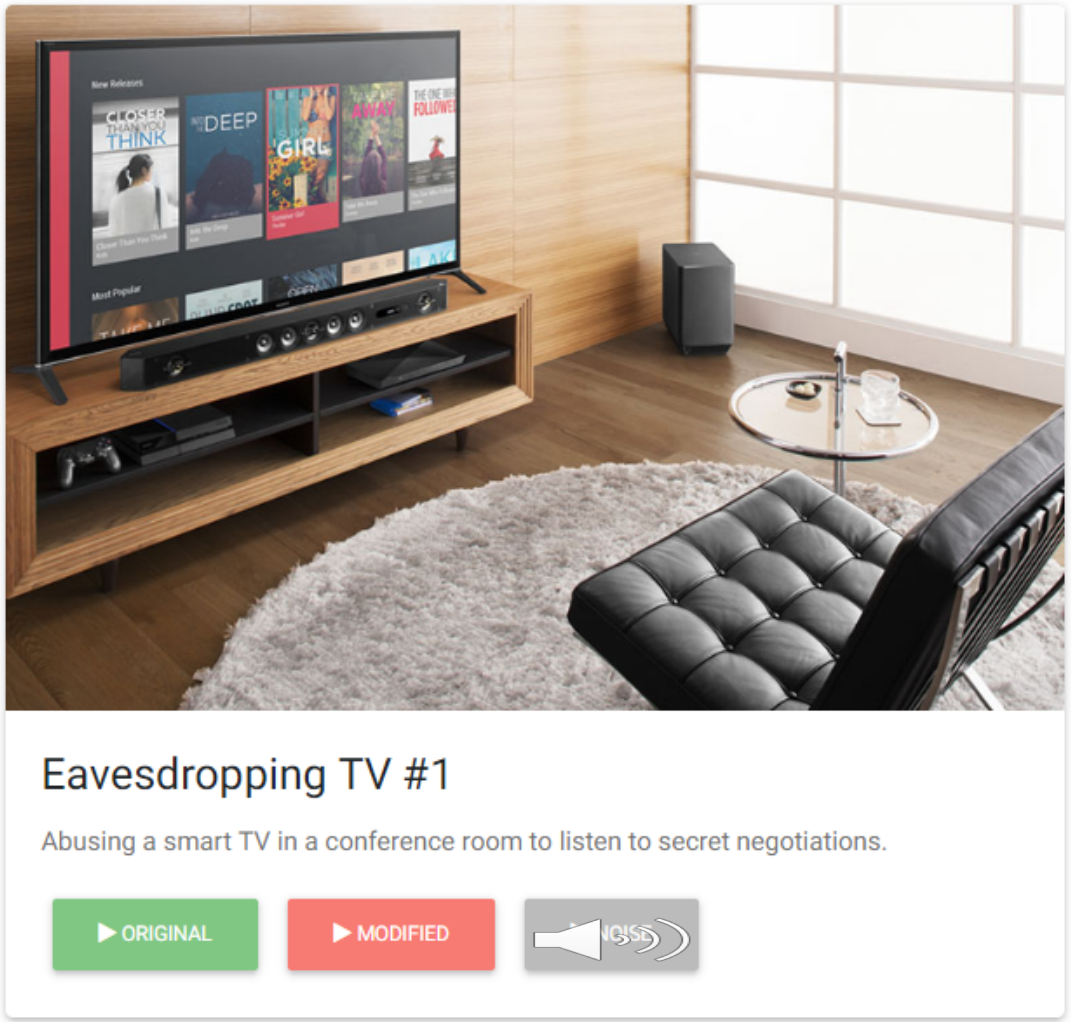
Recognition:

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



Example 1

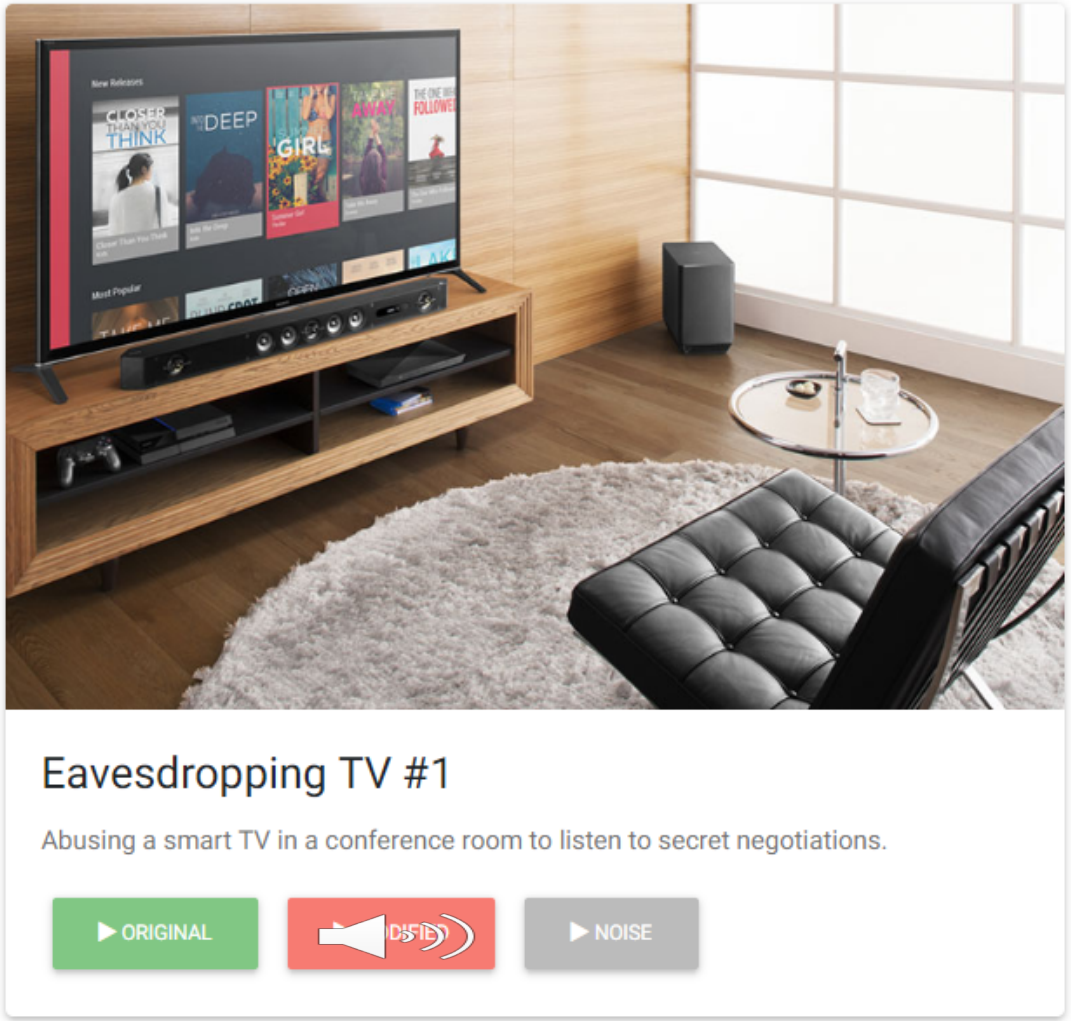
Recognition:



Example 1

Recognition:

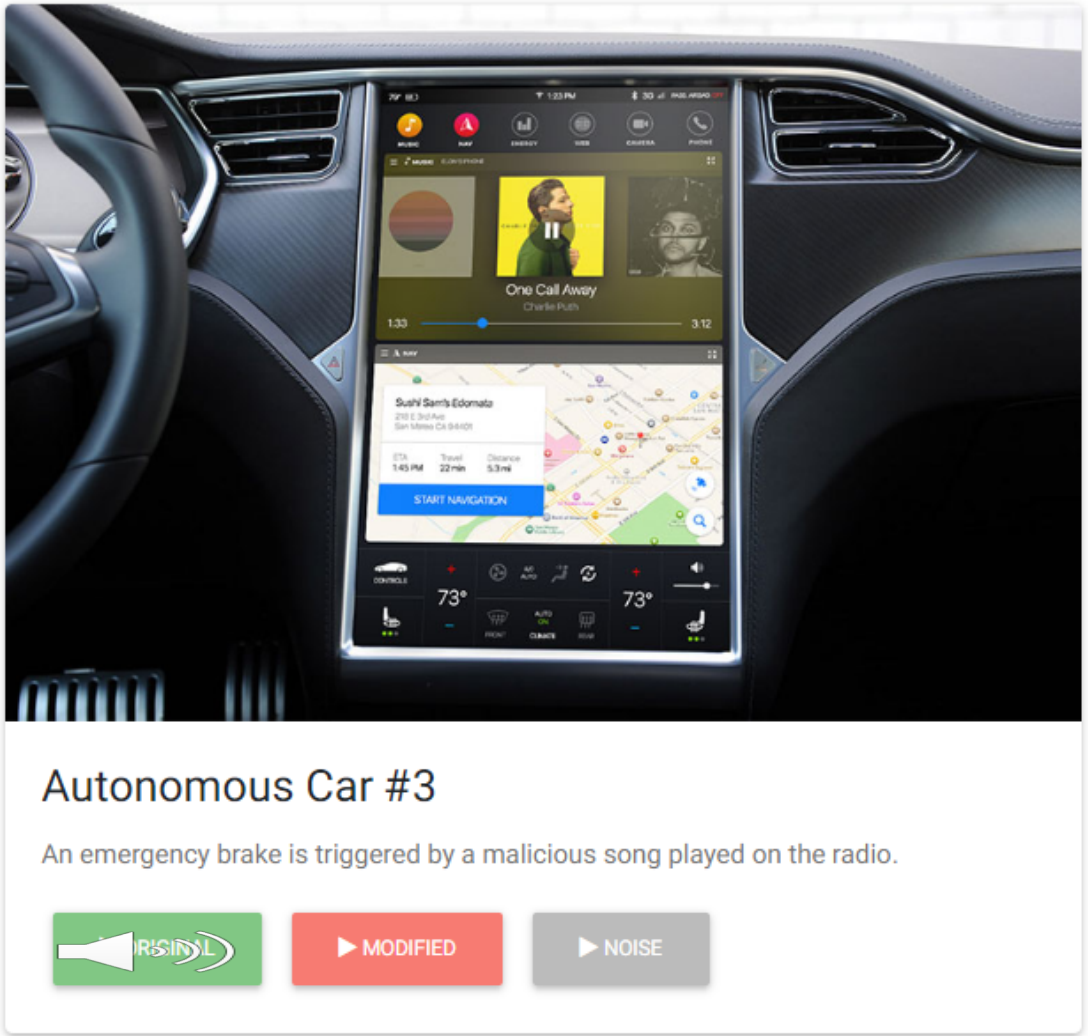
DEACTIVATE SECURITY CAMERA AND UNLOCK FRONT DOOR



Example 2

Recognition:

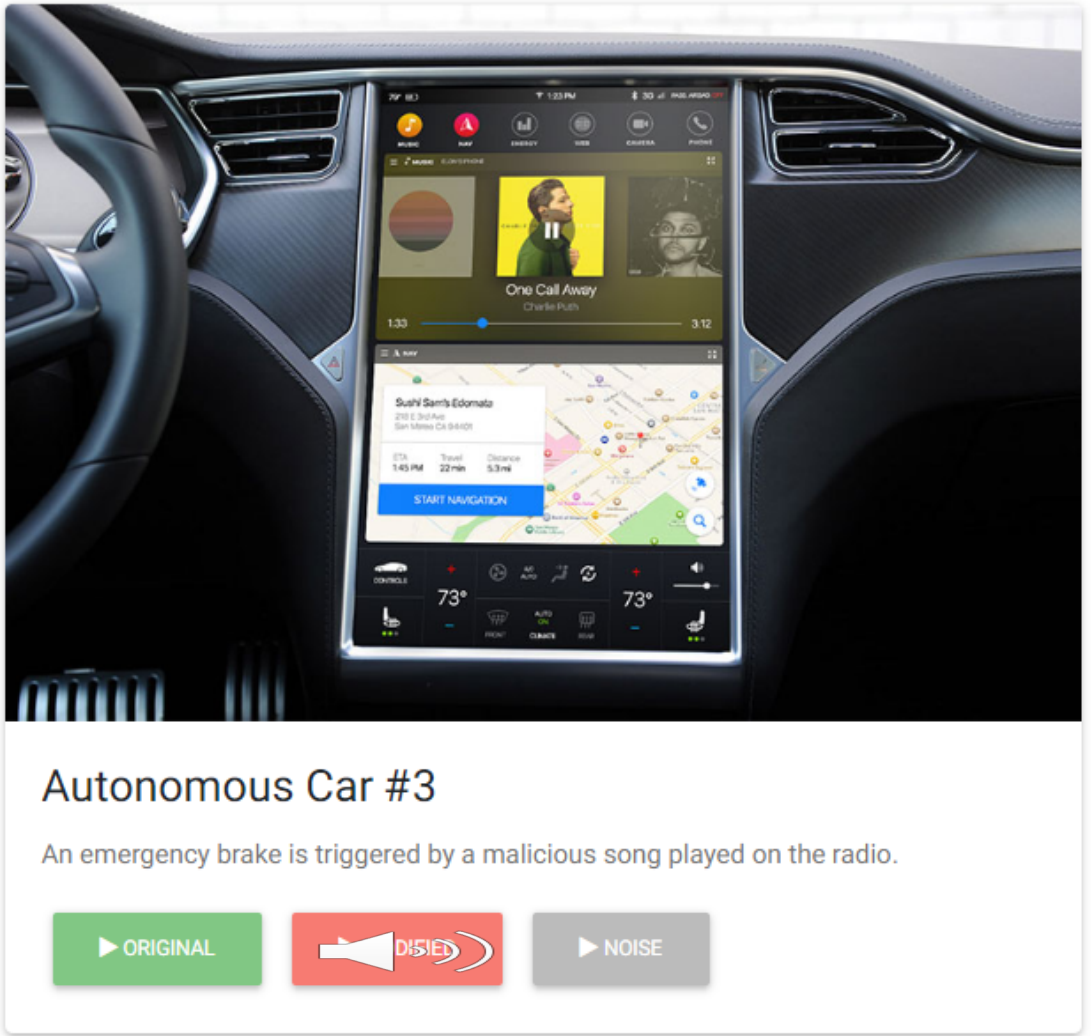
ALAN A NINE
MONTH
UNCERTAIN



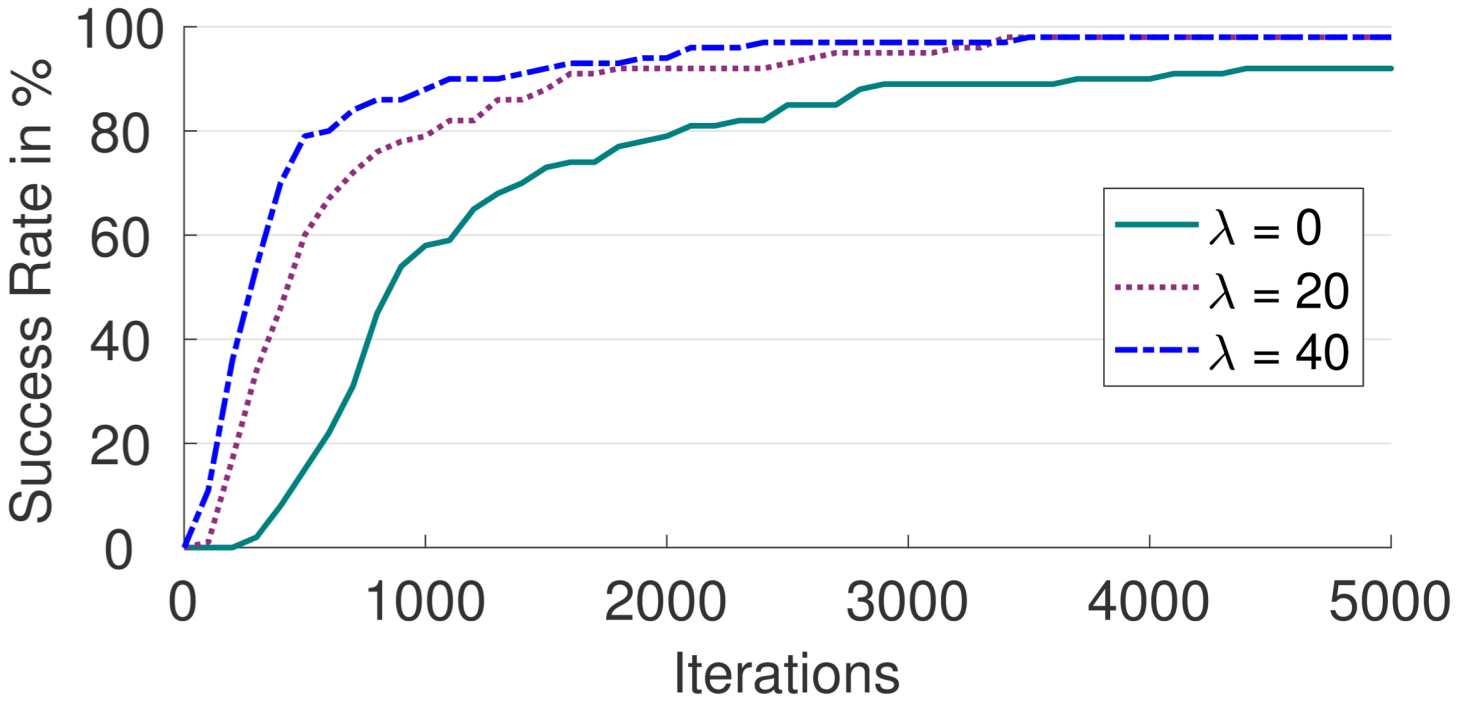
Example 2

Recognition:

ACTIVATE
EMERGENCY
BREAK AND LOCK
ALL DOORS



Performance Analysis



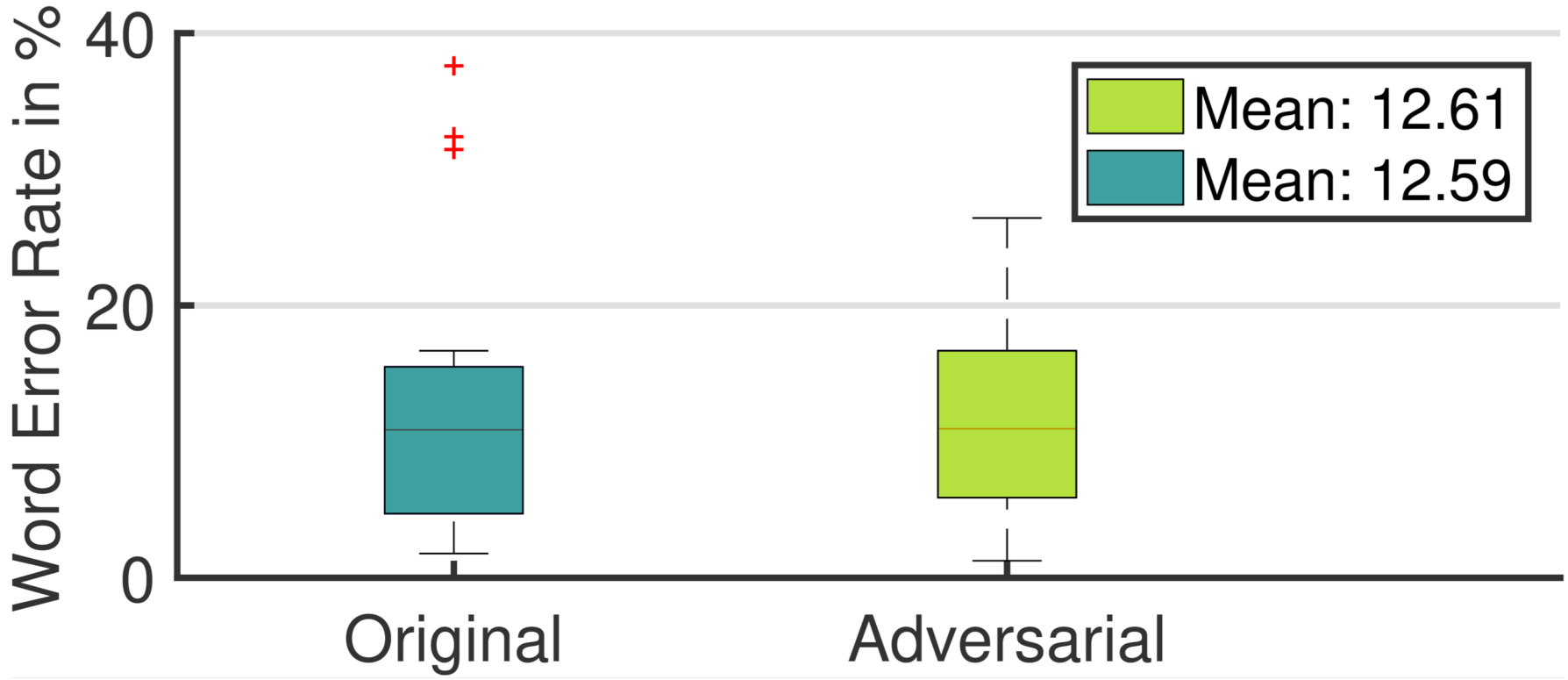
- Only utterances with no missclassifications are counted as success.
- λ describes how much the noise is allowed to exceed the hearing thresholds.

- Comparison with CommanderSong (Yuan et al. USENIX Sec. '18):

	None	λ = 40	λ = 20	λ = 0	CommanderSong
SNR in dB	15.88	17.93	**21.76**	19.38	15.32

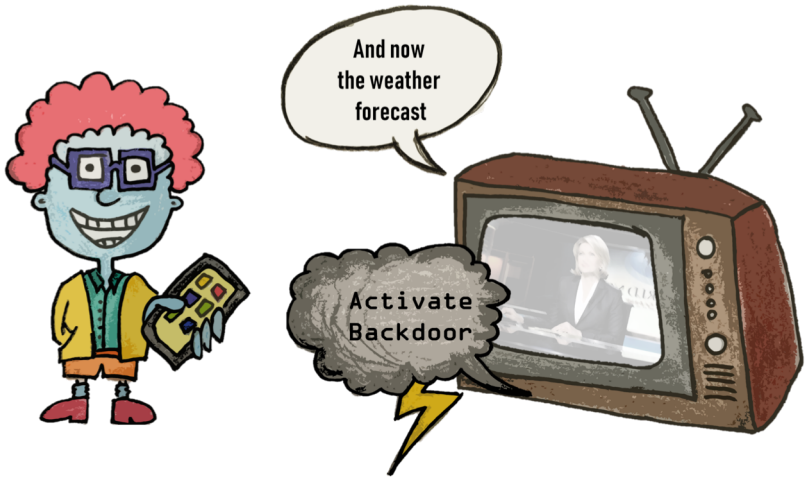
Listening Test - Transcription

- Setup:
  - 22 participants
  - 21 audio examples, with randomly chosen conditions

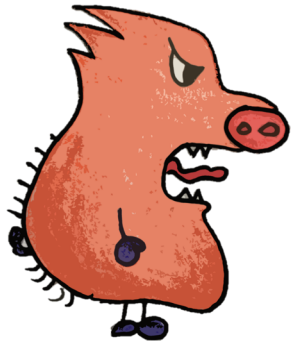


Takeaways

- Adversarial examples for an augmented DNN-HMM hybrid automatic speech recognition are possible
- The added noise can be shaped to remain mostly the hearing thresholds
- The attack works with different kinds of audio content, such as speech, music, or even bird sounds.



Thank You!



Website:

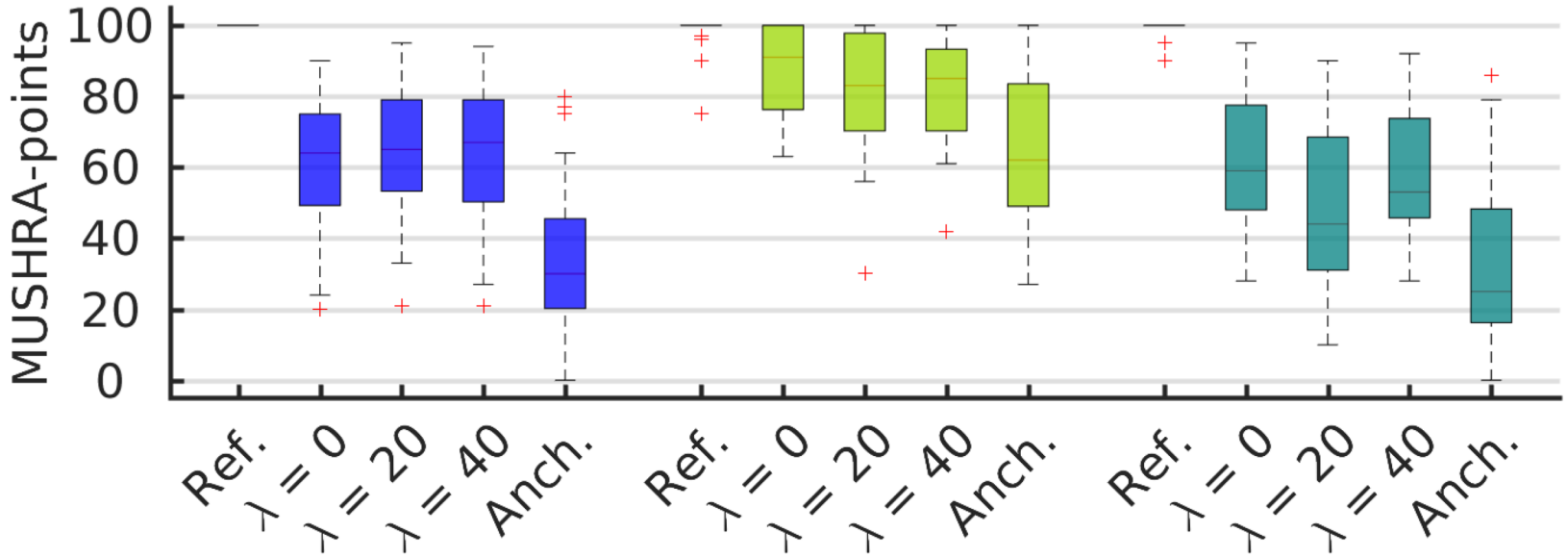
adversarial-attacks.net

Code:

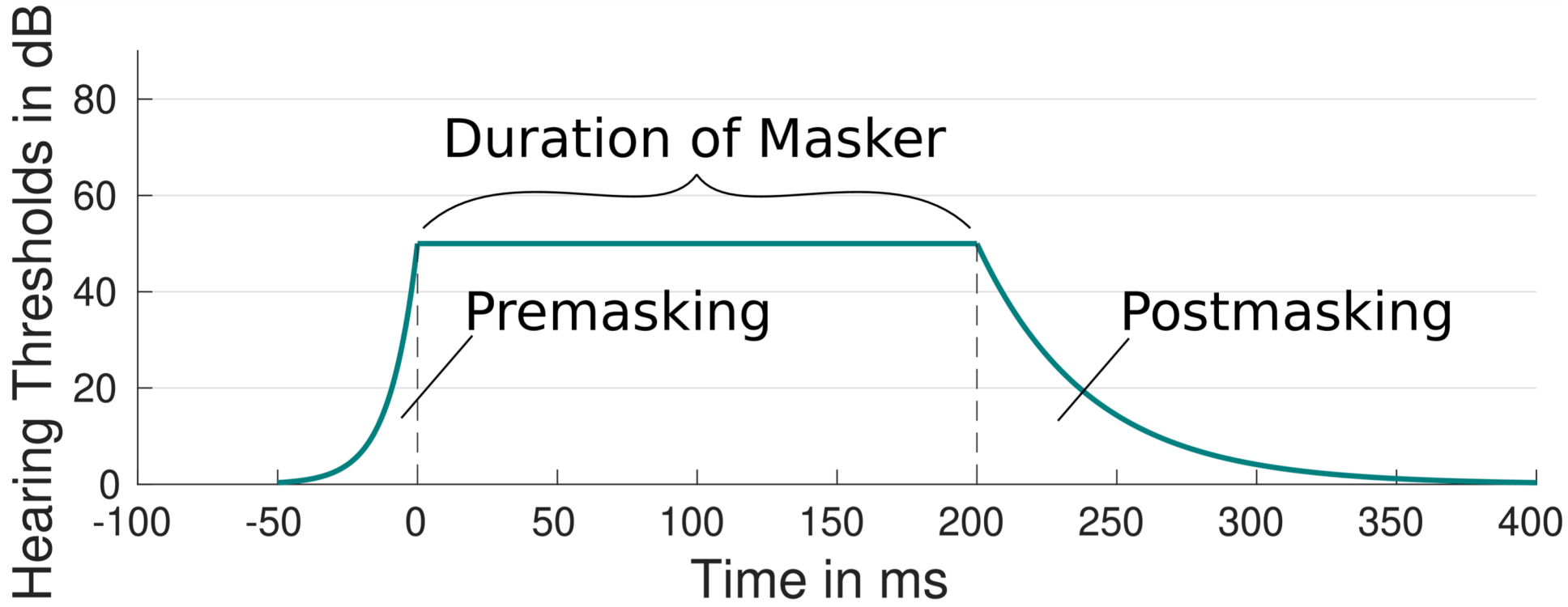
github.com/rub-ksv/adversarialattacks

Listening Test - MUSHRA

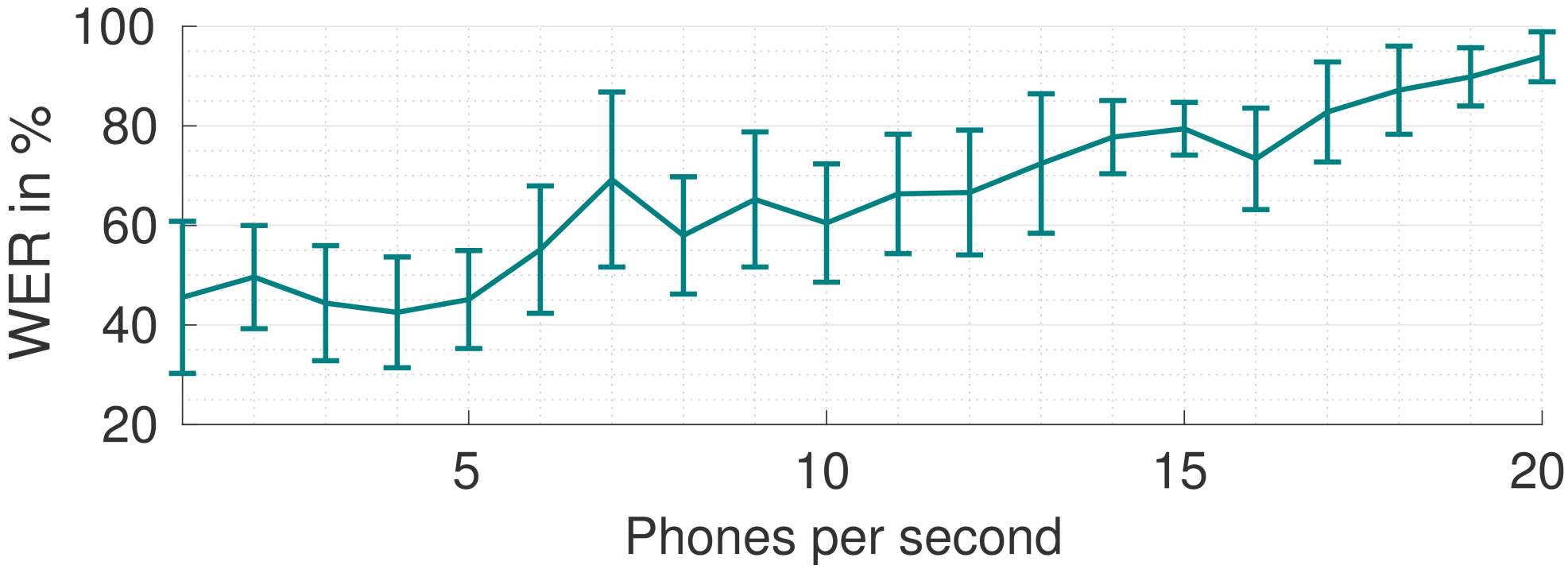
- Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)
- Listening test, where the quality of multiple audio signals is rated in a comparison test



Psychoacoustics – Temporal Masking



Phone Rate Evaluation

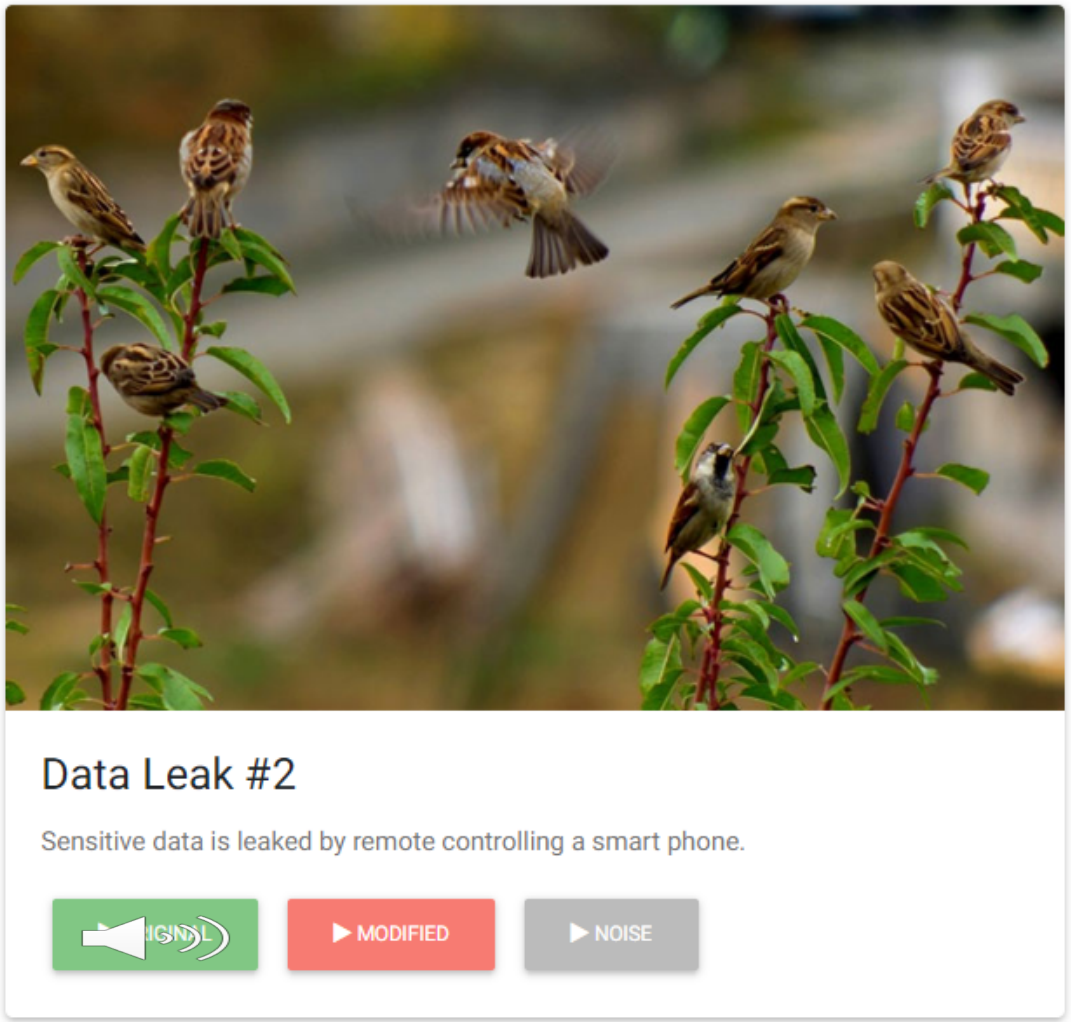


Word Error Rate (WER): Calculated via Levenshtein distance

Example 3

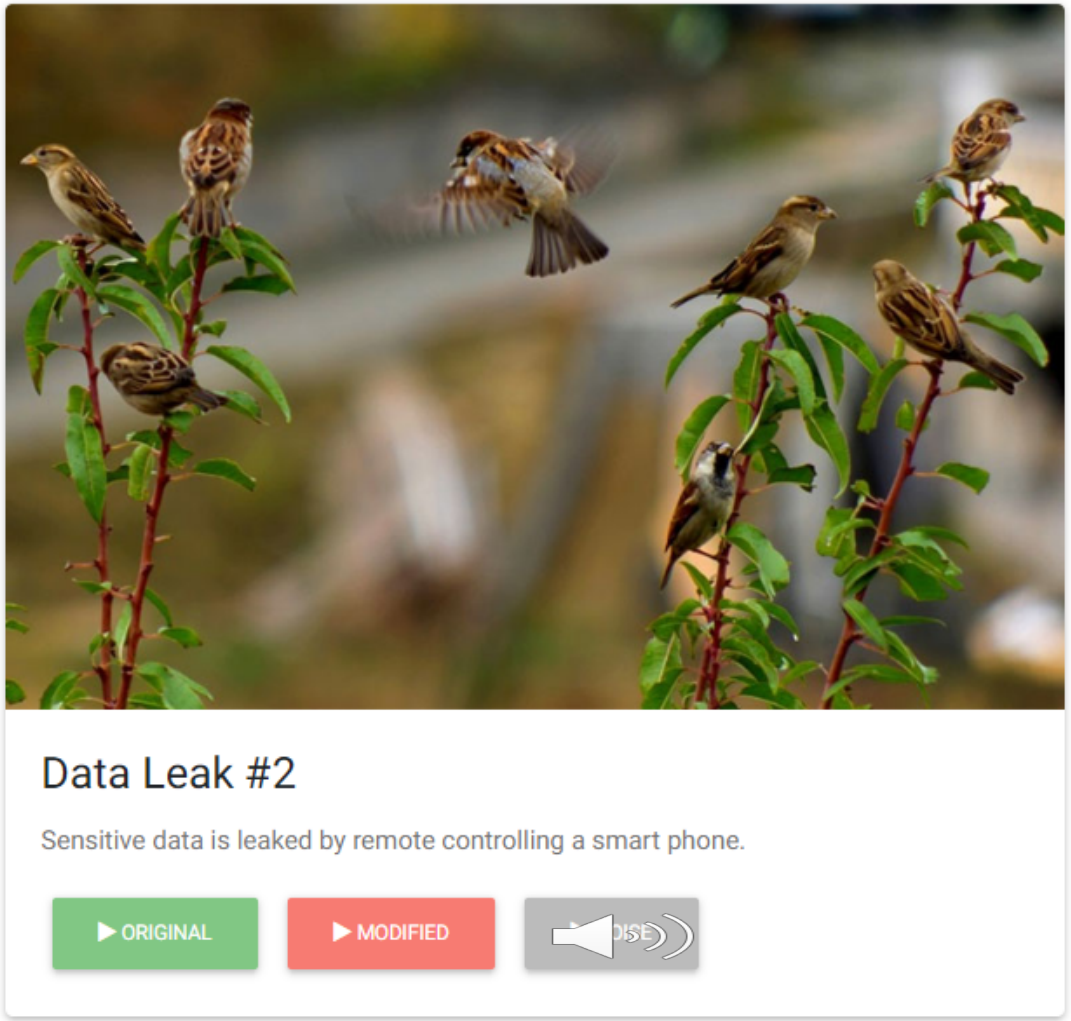
Recognition:

JUDGE FISH



Example 3

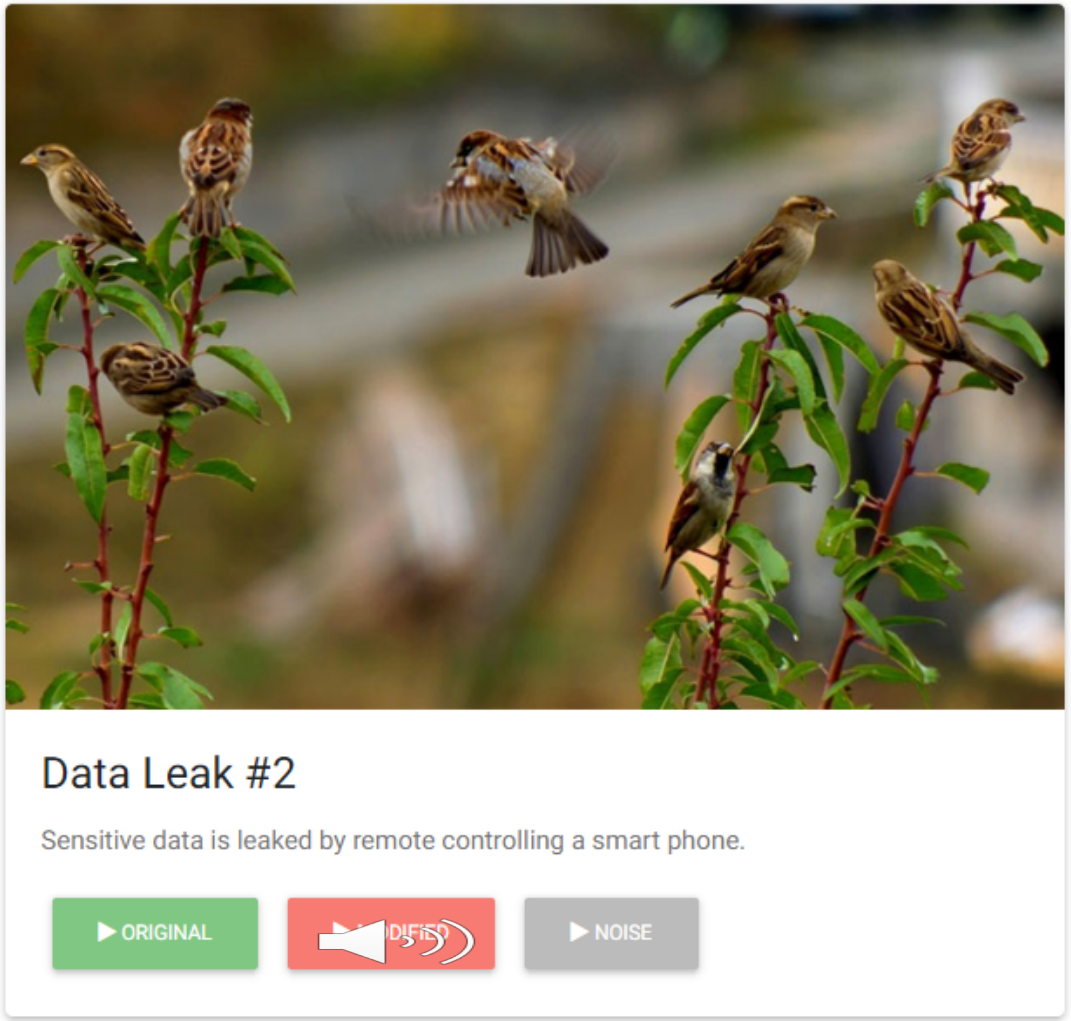
Recognition:



Example 3

Recognition:

VISIT EVIL DOT NET AND INSTALL THE BACKDOOR



Example 2

Recognition: