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 $x$: 

Output $y$: 

$$y = F(x)$$
Adversarial Machine Learning

Input $x + \delta$:

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

Output $y$:

Airliner
Adversarial Machine Learning

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\textsuperscript{[1]} toolkit

---

Automatic Speech Recognition – Feature Extraction

- The feature extraction calculates features in the frequency domain.
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

2. Gradient descent

3. Hearing thresholds

raw audio

"HELLO DARKNESS
MY OLD FRIEND"

original audio

HMM

target transcription

target

$y'$

$pseudo$-posterior

$L(y, y')$

calculate hearing
thresholds

$\alpha$

$\nabla x$
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

1. Framing and Window Function
2. Discrete Fourier Transform
3. Magnitude
4. Logarithm
Psychoacoustics – Frequency Masking

Masking Tone at 1 kHz

Hearing Thresholds in dB

Frequency in kHz

0 1 2 5 10 20

0 20 40 60 80

0.02 0.05 0.1 0.2 0.5
Psychoacoustics – Frequency Masking

Masking Tone at 1 kHz
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 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 UNITED AIRLINES INCLUDING PLANES GATES FACILITIES AND LANDING RIGHTS
Example 1

Recognition:

Eavesdropping TV #1
Abusing a smart TV in a conference room to listen to secret negotiations.

[Buttons: ORIGINAL, MODIFIED, MUTE]
Example 1

Recognition:
DEACTIVATE SECURITY CAMERA AND UNLOCK FRONT DOOR.
Example 2

Recognition:
ALAN A NINE MONTH UNCERTAIN

Autonomous Car #3
An emergency brake is triggered by a malicious song played on the radio.
Example 2

Recognition:
ACTIVATE
EMERGENCY
BREAK AND LOCK
ALL DOORS

Autonomous Car #3
An emergency brake is triggered by a malicious song played on the radio.
Performance Analysis

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

<table>
<thead>
<tr>
<th>SNR in dB</th>
<th>None</th>
<th>$\lambda = 40$</th>
<th>$\lambda = 20$</th>
<th>$\lambda = 0$</th>
<th>CommanderSong</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15.88</td>
<td>17.93</td>
<td>21.76</td>
<td>19.38</td>
<td>15.32</td>
</tr>
</tbody>
</table>

- Only utterances with no missclassifications are counted as success.
- $\lambda$ describes how much the noise is allowed to exceed the hearing thresholds.
Listening Test - Transcription

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

![Graph showing word error rate in % for Original and Adversarial conditions. The graph indicates that the mean word error rate for the Original condition is 12.59%, while for the Adversarial condition it is 12.61%.]
Takeaways

- Adversarial examples for an augmented DNN-HMM hybrid automatic speech recognition are possible.
- The added noise can be shaped to remain mostly within 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

![Diagram showing temporal masking with duration of masker, premasking, and postmasking.](image)

- **Duration of Masker**
- **Premasking**
- **Postmasking**

The graph illustrates the change in hearing thresholds in dB over time in milliseconds (ms).
Phone Rate Evaluation

Word Error Rate (WER): Calculated via Levenshtein distance
Example 3

Recognition:
JUDGE FISH
Example 3

Recognition:

Data Leak #2
Sensitive data is leaked by remote controlling a smart phone.
Example 3

Recognition:
VISIT EVIL DOT NET AND INSTALL THE BACKDOOR
Example 2

Recognition:

Autonomous Car #3
An emergency brake is triggered by a malicious song played on the radio.