Attacks as Defenses: Designing Robust Audio CAPTCHAs Using Attacks on Automatic Speech Recognition Systems

Hadi Abdullah, Aditya Karlekar, Saurabh Prasad, Muhammad Sajidur Rahman, Logan Blue, Luke A. Bauer, Vincent Bindschaedler, Patrick Traynor

NDSS Symposium 2023
March 1, 2023
Understanding Can be Dangerous
Audio CAPTCHAs

*Garbage*
Audio CAPTCHAs
Goal

Design *high quality* audio CAPTCHAs that are *robust* to ASRs based on the differences between how humans and machines understand audio.
Criteria

• Human Intelligibility
• ASR UnIntelligibility
• Adaptive Adversary
• Misuse Detection
Criteria

- Human Intelligibility
- ASR UnIntelligibility
- Adaptive Adversary
- Misuse Detection

Abc123
Criteria

- Human Intelligibility
- ASR UnIntelligibility
- Adaptive Adversary
- Misuse Detection
Criteria

- Human Intelligibility
- ASR UnIntelligibility
- Adaptive Adversary
- Misuse Detection
Criteria

• Human Intelligibility
• ASR UnIntelligibility
• Adaptive Adversary
• Misuse Detection
Evaluating Current Methods

- Human Intelligibility
- ASR UnIntelligibility
- Adaptive Adversary
- Misuse Detection

| Reference          |
|--------------------|
| Taori et al. [92]  |
| M. Azalnot et al. [25] |
| HVC (2) [39]       |
| Cocaine Noodles [94] |
| Dolphin Attack [102] |
| Light Commands [89] |
| Roy et al. [72]    |
| HVC (1) [39]       |
| CW [40]            |
| Houdini [45]       |
| Schonherr et al. [79] |
| Kreuk et al. [57]  |
| Qin et al. [69]    |
| Yakura et al. [99] |
| Commander Song [101] |
| Devil’s Whisper [42] |
| Abdoli et al. [18] |
| P-PGD [22]         |
| Kenansville Attack [21] |
| Abdulllah et al. [19] |
Human Intelligibility

- ASR UnIntelligibility
- Adaptive Adversary
- Misuse Detection
### ASR UnIntelligibility

- Human Intelligibility
- ASR UnIntelligibility
- Adaptive Adversary
- Misuse Detection

| Method                          | Intelligible | Opt  |
|--------------------------------|--------------|------|
| CW [40]                        | ✔            | Opt  |
| Houdini [45]                   | ✔            | Opt  |
| Schonherr et al. [79]          | ✔            | Intelligible |
| Kreuk et al. [57]              | ✔            | Intelligible |
| Qin et al. [69]                | ✔            | Intelligible |
| Yakura et al. [99]             | ✔            | Intelligible |
| Commander Song [101]           | ✔            | Intelligible |
| Devil’s Whisper [42]           | ✔            | Intelligible |
| Abdoli et al. [18]             | ✔            | Intelligible |
| P-PGD [22]                     | ✔            | Intelligible |
| Kenansville Attack [21]        | ✔            | Intelligible |

*Sig Proc*
ASR UnIntelligibility

- Human Intelligibility
- ASR UnIntelligibility
- Adaptive Adversary
- Misuse Detection

- CW [40]
- Houdini [45]
- Schonherr et al. [79]
- Kreuk et al. [57]
- Qin et al. [69]
- Yakura et al. [99]
- Commander Song [101]
- Devil’s Whisper [42]
- Abdoli et al. [18]
- P-PGD [22]
ASR UnIntelligibility

- Human Intelligibility
- ASR UnIntelligibility
- Adaptive Adversary
- Misuse Detection

| Method                        | Intelligible | Opt   |
|-------------------------------|--------------|-------|
| CW [40]                       | ✓            | Intelligible | Opt   |
| Houdini [45]                  | ✓            | Intelligible | Opt   |
| Schonherr et al. [79]         | ✓            | Intelligible | Opt   |
| Kreuk et al. [57]             | ✓            | Intelligible | Opt   |
| Qin et al. [69]               | ✓            | Intelligible | Opt   |
| Yakura et al. [99]            | ✓            | Intelligible | Opt   |
| Commander Song [101]          | ✓            | Intelligible | Opt   |
| Devil's Whisper [42]          | ✓            | Intelligible | Opt   |
| Abdoli et al. [18]            | ✓            | Intelligible | Opt   |
| P-PGD [22]                    | ✓            | Intelligible | Opt   |
| Kenansville Attack [21]       | ✓            | Intelligible | Sig Proc |
Adaptive Adversary

- Kenansville Attack [21]
- Add Gaussian Noise

“123”

- Human Intelligibility
- ASR UnIntelligibility
- Adaptive Adversary
- Misuse Detection
Key Takeaways:

- Human Intelligibility
- ASR UnIntelligibility
- Adaptive Adversary Misuse Detection
New Attack Strategy

Simplified ASR Pipeline

- Feature Extraction
- CNN
- RNN
- char
- Dictionary
- Language Model
Feature Extraction

\[ F_k = \sum_{n=0}^{N-1} s_n \left( \cos \left( \frac{\pi}{N} (n + \frac{1}{2})k \right) - i \cdot \sin \left( \frac{\pi}{N} (n + \frac{1}{2})k \right) \right) \]

\[ F_k = |F_k| \]

\[ m_k = 2595 \log_{10} \left( 1 + \frac{|F_k|}{700} \right) \]

\[ F_k = \sum_{n=0}^{N-1} s_n \cos \left( \frac{(2n + 1)k\pi}{2N} \right) \]
Feature Extraction Ignores Psychoacoustics

\[ F_k = \sum_{n=0}^{N-1} s_n \left( \cos\left(\frac{\pi}{N} (n + \frac{1}{2})k\right) - i \cdot \sin\left(\frac{\pi}{N} (n + \frac{1}{2})k\right) \right) \]

\[ F_k = |F_k| \]

\[ s_k = 2595 \log_{10} \left( 1 + \frac{F_k}{700} \right) \]

\[ F_k = \sum_{n=0}^{N-1} s_n \cos\left(\frac{(2n + 1)k\pi}{2N}\right) \]

1. Lossy
2. Psychoacoustics

Psychoacoustics far more complex:
1. Frequency masking.
2. Cocktail-party effect.
3. Ignoring low intensity frequencies.
4. ...etc
**Intuition**

**Human Ear: Formant Dependence**
1. Can understand modified formants.

**ASRs :(**

**Clipping formants:**
1. Maintain audio quality for the human ear.
2. Force ASRs to output Empty String.
Algorithm

Original Audio Sample → YeeHaw Junction Algorithm → Perturbed Audio Sample → ASR

*Empty String*
YeeHaw Junction is better than reCaptcha

| reCAPTCHA | YeeHaw Junction |
|------------|-----------------|
| Vulnerability Against Bots | ✔️ Human Intelligibility |
| User Error Rate (via User Study) | ✔️ ASR UnIntelligibility |
|   | • Adaptive Adversary |
|   | • Misuse Detection |
## Final Takeaways

|                | Optimization Attacks | Signal Processing Attacks | Yeehaw Junction |
|----------------|----------------------|---------------------------|-----------------|
| **Human Intelligibility** | ✓                    | ✓                          | ✓               |
| **ASR Unintelligibility**   | ✗                    | ✓                          | ✓               |
| **Adaptive Adversary**      | ✗                    | ✗                          | ✓               |
| **Misuse Detection**        |                       |                           | ✓               |
Final Takeaways

- We design Yeehaw Junction, a principles technique to make Audio CAPTCHAs
- ASR transcribes at a rate of 0.004%
- Improved audio quality compared to existing techniques

Hadi Abdullah
hadiabdullah.github.io
habdulla@visa.com

Logan Blue
lblue.us
bluel@ufl.edu