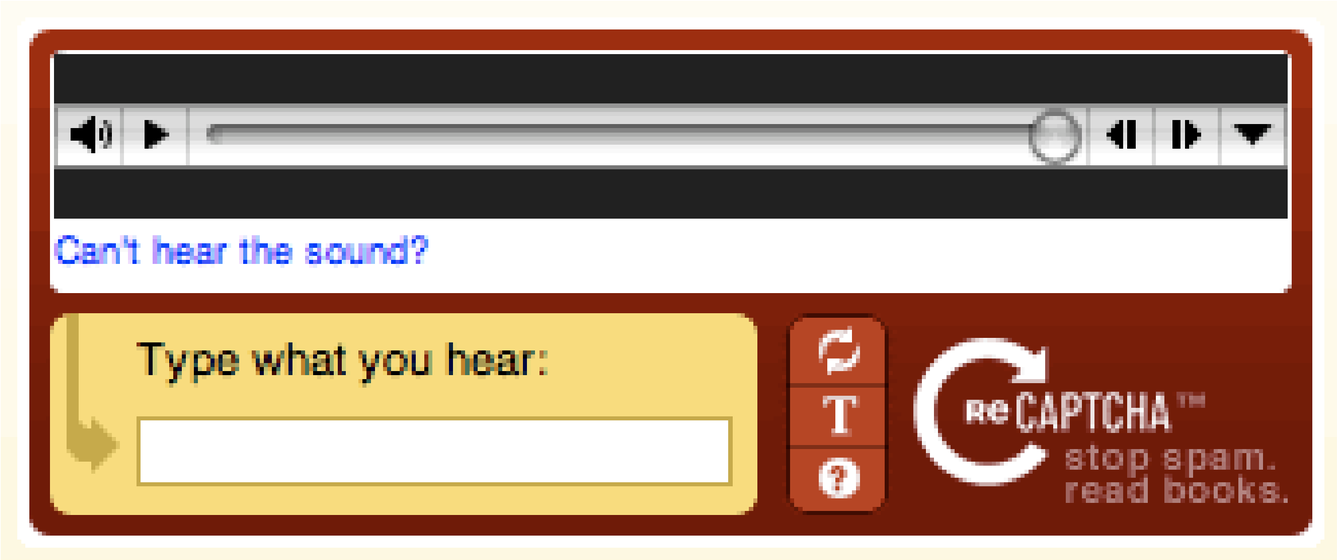
Attacks as Defenses: Designing Robust Audio CAPTCHAs Using Attacks on Automatic Speech Recognition Systems

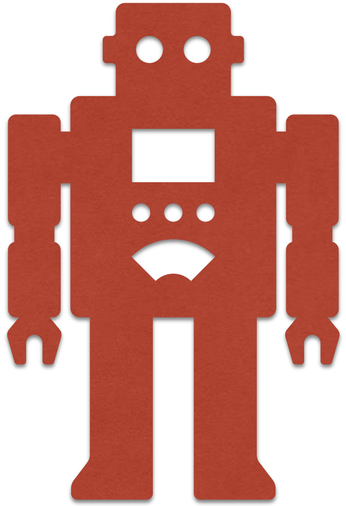
Hadi Abdullah, Aditya Karlekar, Saurabh Prasad,
Muhammad Sajidur Rahman, **Logan Blue**, Luke A. Bauer,
Vincent Bindschaedler, Patrick Traynor

NDSS Symposium 2023
March 1, 2023

Understanding Can be Dangerous

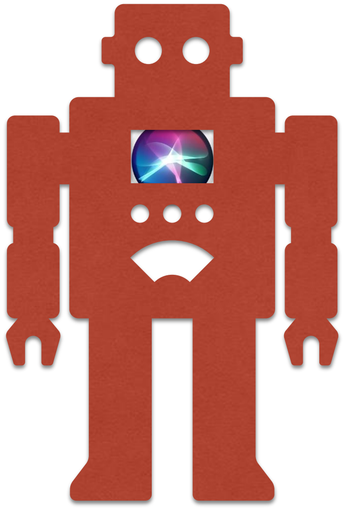
Audio CAPTCHAs





Garbage

Audio CAPTCHAs



abc123

Goal

Design *high quality* audio CAPTCHAs

that are *robust* to ASRs based on the differences between

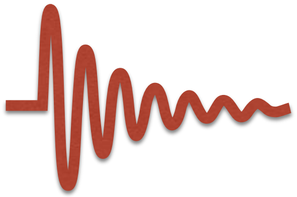
how humans and machines understand audio.

Criteria

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

Criteria

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



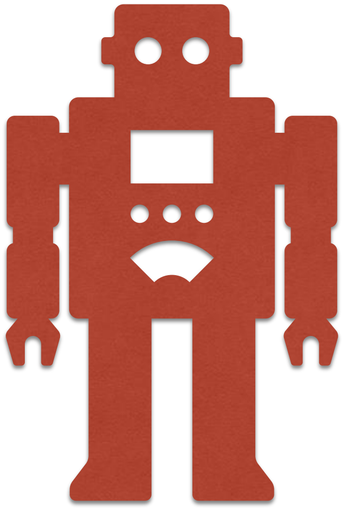
Abc123

Criteria

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

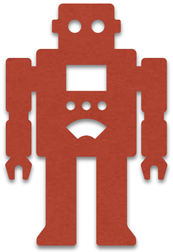
Criteria

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



Criteria

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



Evaluating Current Methods

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]
Abdullah et al. [19]

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

Human Intelligibility

	Potential CAPTCHA Use	Audio Quality	Attack Type
Taori et al. [92]	✗	Intelligible	Grad Free
M. Azalnot et al. [25]	✗	Intelligible	Grad Free
HVC (2) [39]	✗	Inaudible	Misc
Cocaine Noodles [94]	✗	Inaudible	Misc
Dolphin Attack [102]	✗	Inaudible	Misc
Light Commands [89]	✗	Inaudible	Misc
Roy et al. [72]	✗	Inaudible	Misc
HVC (1) [39]	✓	Unintellgible	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
Abdullah et al. [19]	✗	Unintellgible	Sig Proc

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

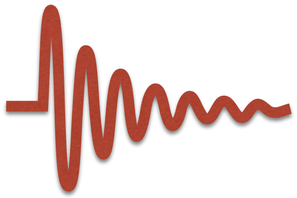
ASR UnIntelligibility

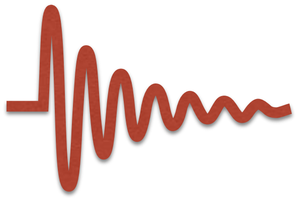
CW [40]	✓	Intelligible	Opt
Houdini [45]	✓	Intelligible	Opt
Schonherr et al. [79]	✓	Intelligible	Opt
Kreuk et al. [57]	✓	Intelligible	Opt
Qin et al. [69]	✓	Intelligible	Opt
Yakura et al. [99]	✓	Intelligible	Opt
Commander Song [101]	✓	Intelligible	Opt
Devil's Whisper [42]	✓	Intelligible	Opt
Abdoli et al.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

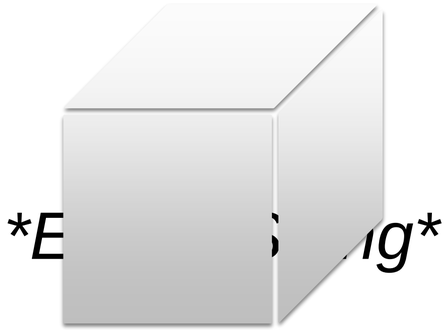
- ✓ Human Intelligibility
- ASR UnIntelligibility
- Adaptive Adversary
- Misuse Detection

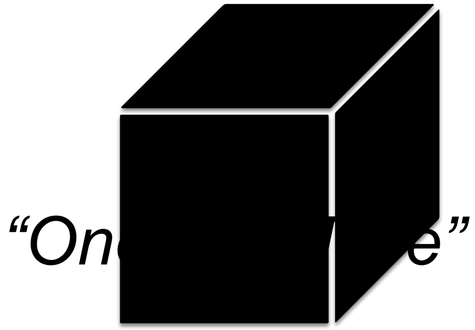
ASR UnIntelligibility

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]









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

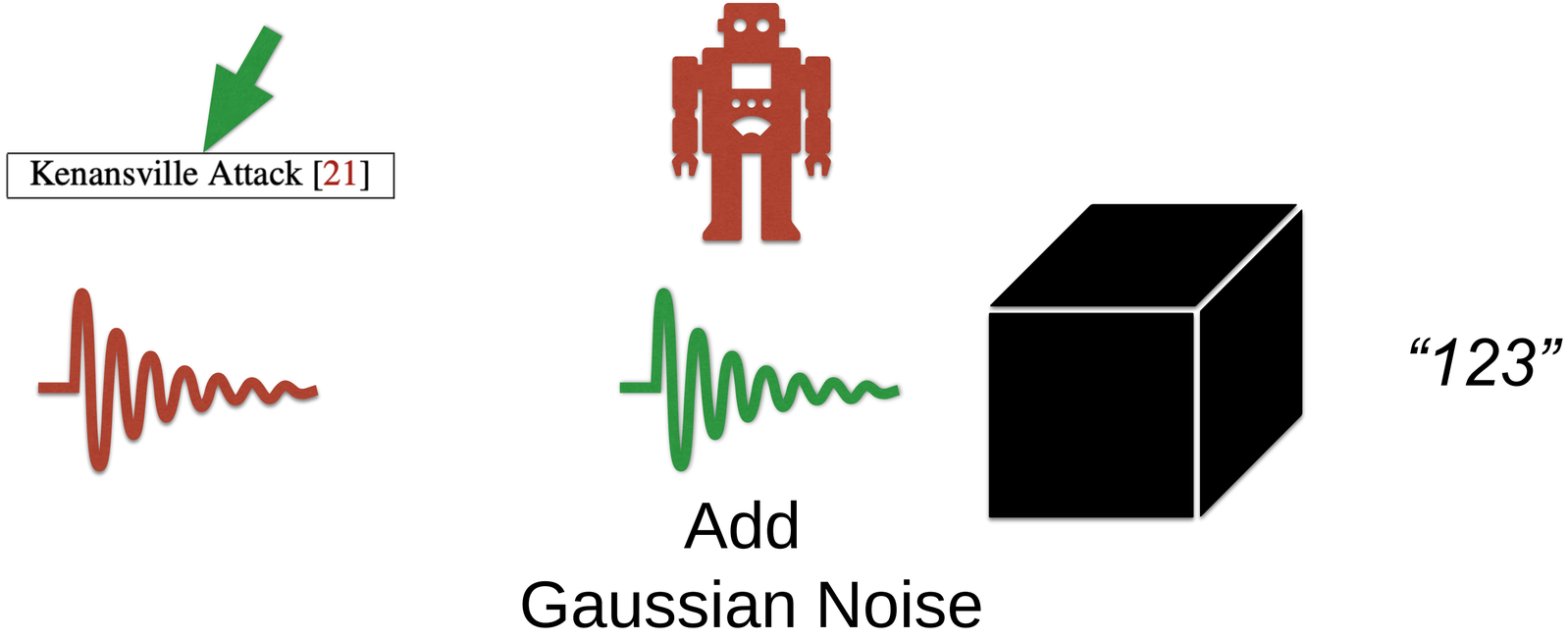
ASR UnIntelligibility

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

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

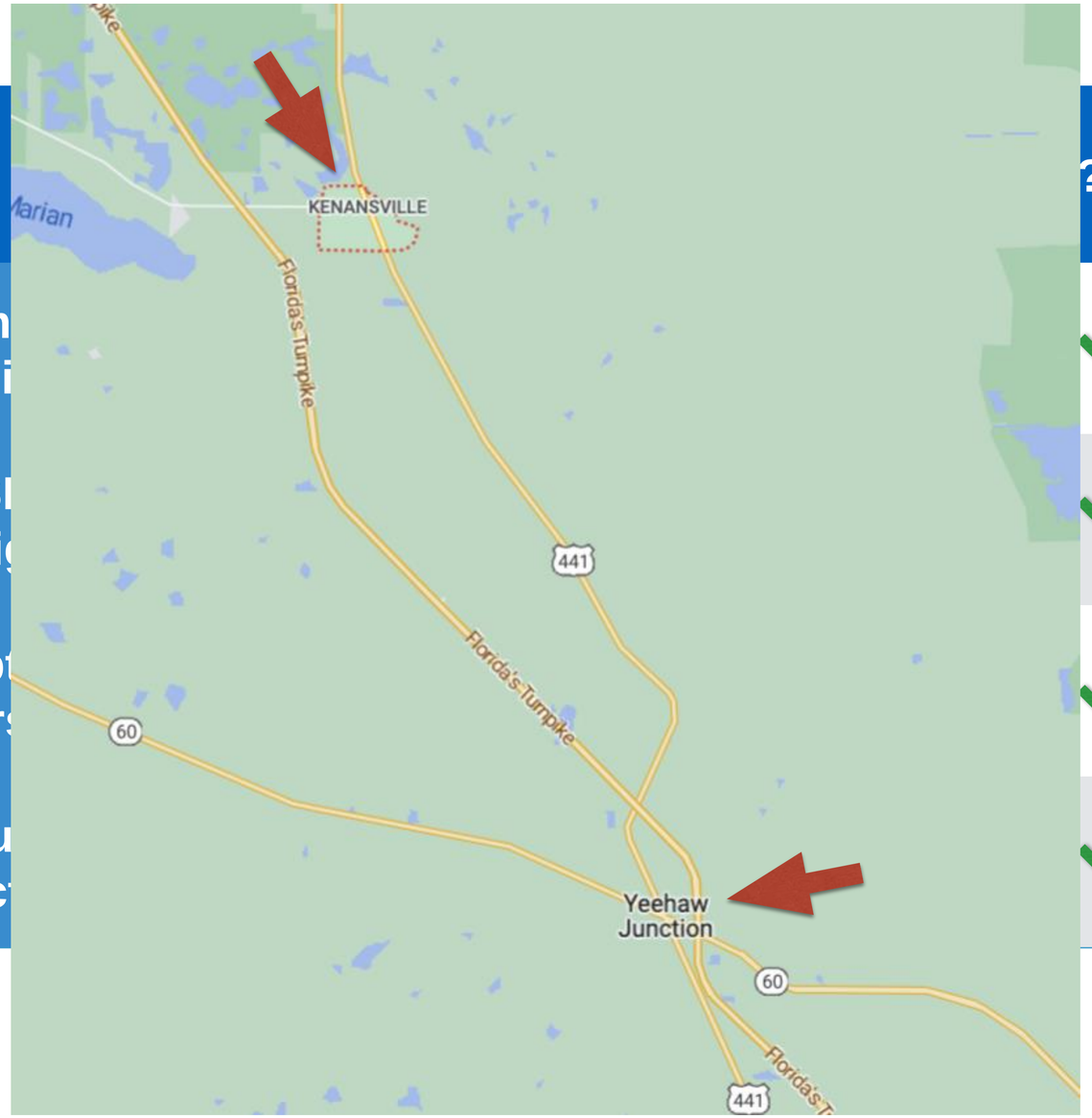
Adaptive Adversary



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

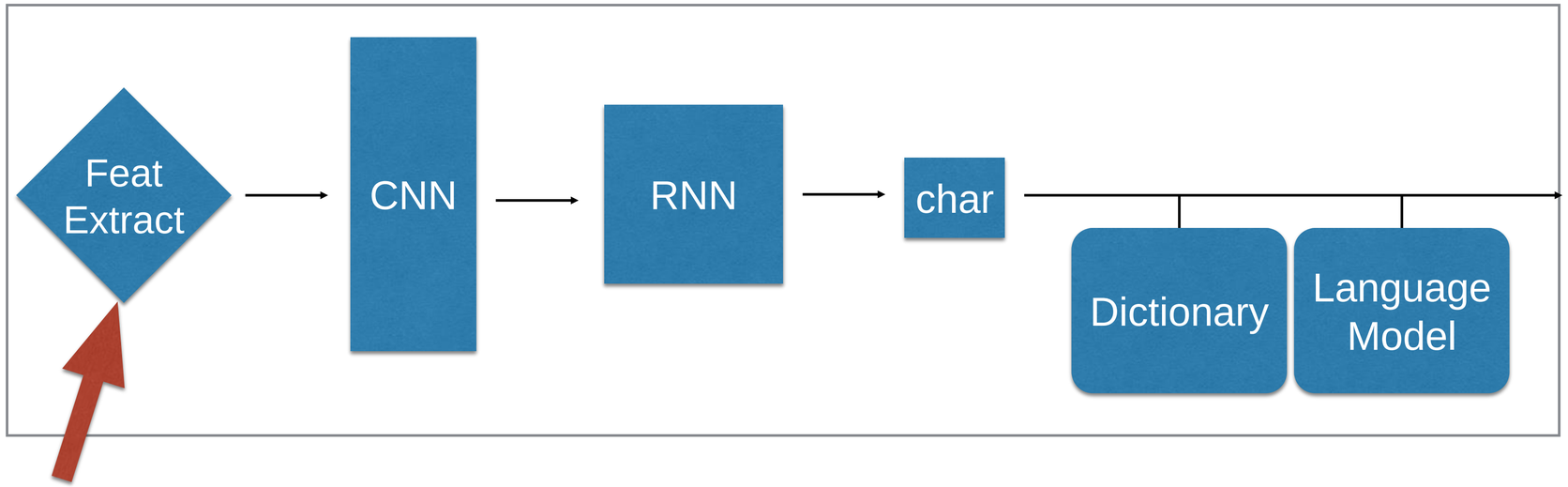
Key Takeaways:

	???
Hum Intelligi	✓
ASI UnIntelli	✓
Adapt Advers	✓
Misu Detec	✓



New Attack Strategy

Simplified ASR Pipeline



Feature Extraction

DCT

Feat Extract

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

$$F_k = |F_k|$$

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

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

Feature Extraction Ignores Psychoacoustics

DFT

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

Mag

$$F_k = |F_k|$$

Mel

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

Scaling

DCT

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

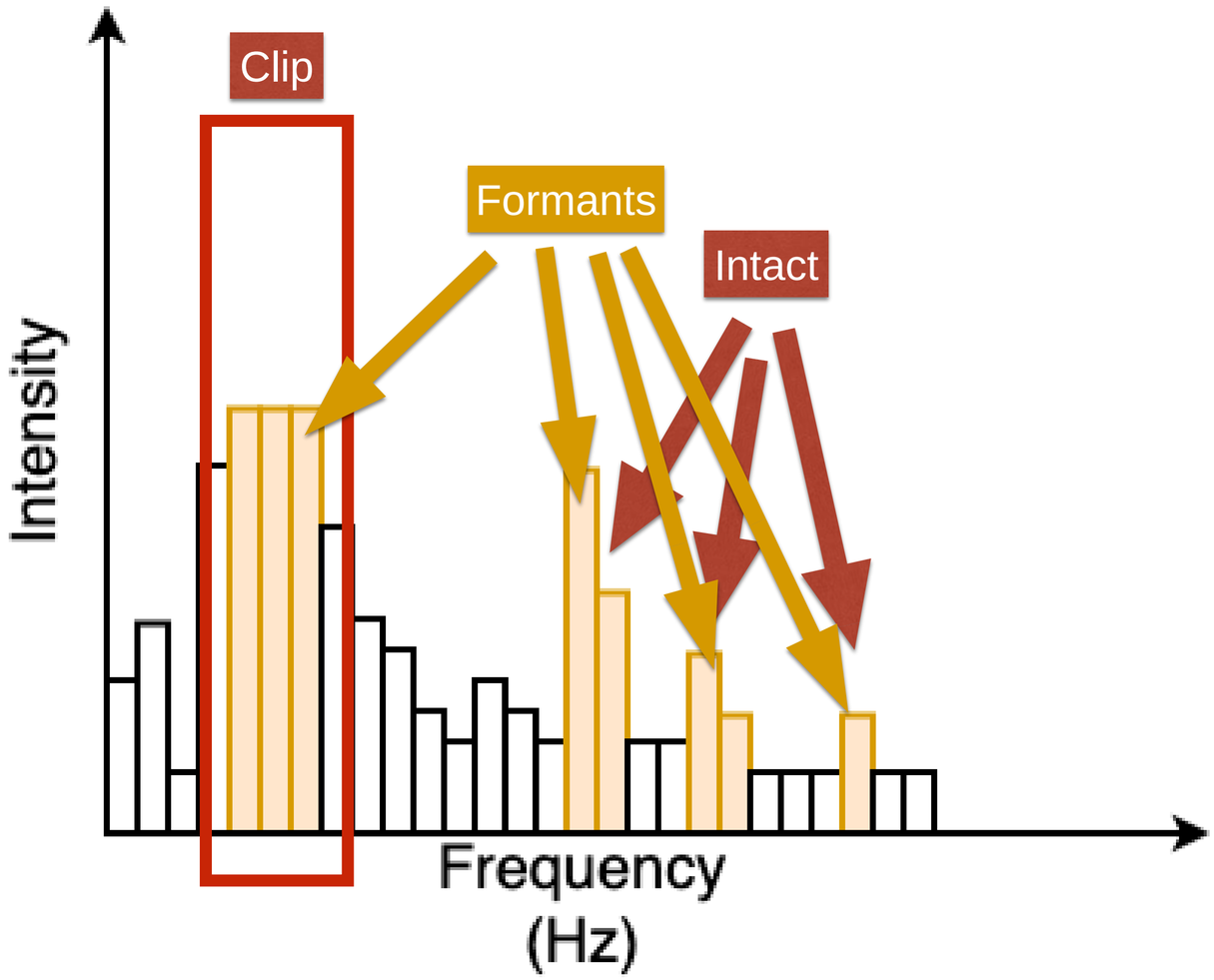
1. Lossy

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

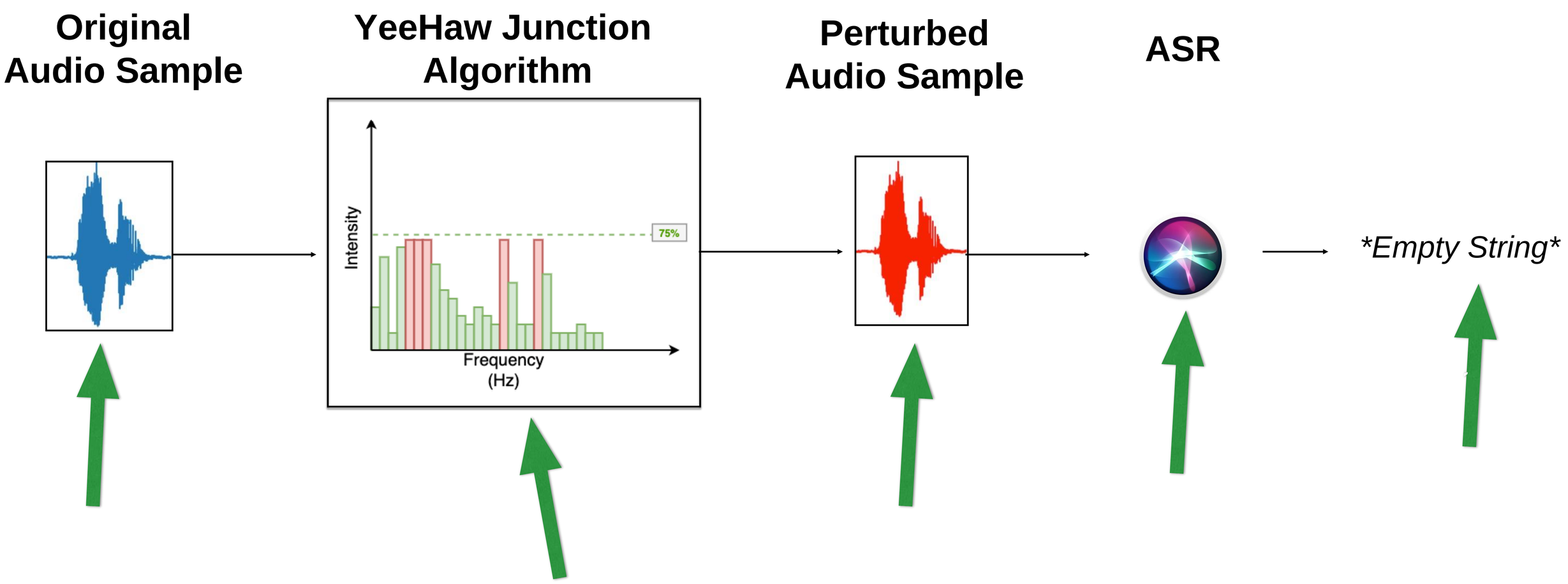
1. Can understand modified formants.

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
Intensity
75%
Frequency
(Hz)
Perturbed
Audio Sample
ASR
Empty String

YeeHaw Junction is better than reCaptcha

	reCAPTCHA	YeeHaw Junction
Vulnerability Against Bots		
User Error Rate (via User Study)		

- ✓ Human Intelligibility
- ✓ 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 Ye make Audio CAPTC
- ASR transcribe
- Improved audi es

lah
thub.io
a.com

e
lblue.us
bluel@ufl.edu