**Introduction**

We generate adversarial examples such that a human perceives the audio as “Yes” while a machine recognizes it as “No”.

- Automatic speech recognition (ASR) is used for various applications: digital assistants, smart-home devices, telephone response.
- Prior work on adversarial attacks focused mainly on image recognition and object detection models.
- Adversarial attacks can potentially disrupt these applications.

**Challenges:**
- Existing gradient-based method of adversarial attacks (e.g. FGSM, DeepFool, Carlini) are not suited to perform adversarial attacks against speech recognition models:
  - They require the recognition pipeline to be differentiable.
  - Typical automatic speech recognition models include steps that compute spectrograms and MFCC features, these operations are not differentiable.
- We propose a novel adversarial attack on ASR based on genetic optimization.
- We do targeted attacks not showcased before.

**Methodology**

```
“Yes” audio clip

Randomly perturb 8 least significant bits

Population of attack clips

attack success?

Yes

Attack clip found

Evaluate fitness score for population members

Select candidate parent clips with high scores

Crossover: Randomly select which part of parent clip should go to child

Parent clip 1

Parent clip 2

child clip

Mutate: randomly perturb 8 least significant bits

add attack clip for next population
```

**Results**

- Conducted human experiment with 23 participants who labeled nearly 1500 successful attack audio clips.
- The effect of adversarial noise on the human perception is negligible.

<table>
<thead>
<tr>
<th>Attack Labeled as Source</th>
<th>Attack Labeled as Target</th>
<th>Attack Labeled as Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>89%</td>
<td>0.8%</td>
<td>9.4%</td>
</tr>
</tbody>
</table>

Table: Human perception of adversarial examples. Results from 1500 human labeling of our adversarial audio clips.

**Code**

**Audio Samples**

**Bibliography**


**Results**

- Conducted human experiment with 23 participants who labeled nearly 1500 successful attack audio clips.
- The effect of adversarial noise on the human perception is negligible.