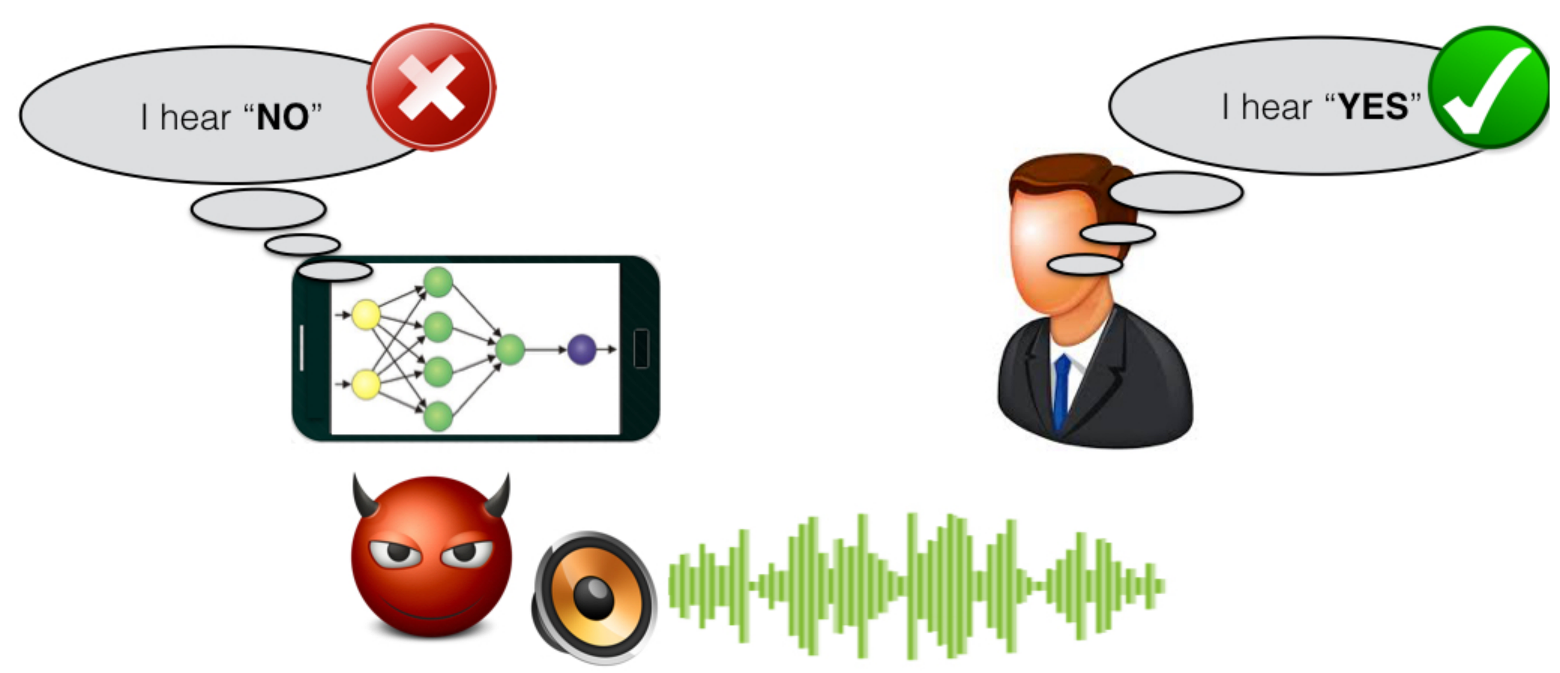
Did you hear that? Adversarial Examples Against Automatic Speech Recognition

Moustafa Alzantot, Bharathan Balaji, Mani Srivastava
University of California, Los Angeles (UCLA)

Introduction



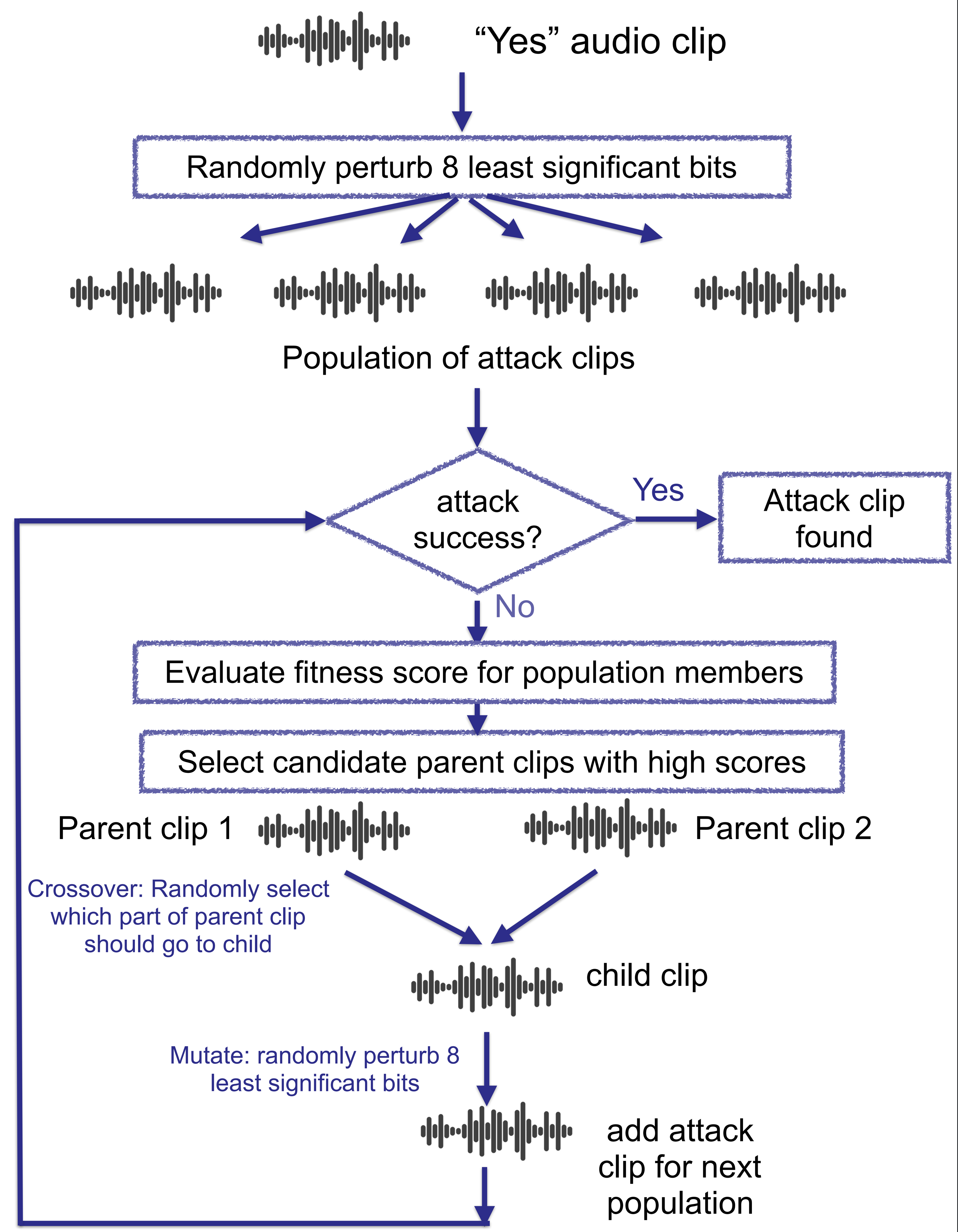
We generate adversarial examples such that a human perceives the audio as "Yes" while a machine recognizes it as "No"

- 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

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



Flowchart: Overview of our genetic algorithm based attack

- Evaluated using Speech Commands dataset.
- 65000 1 second audio files, 10 words
- Perform targeted attacks against 500 random files of each word to every other word label.
  - Generated 4500 output files.
- Average attack success rate = 87%.

Results

	yes	no	up	down	left	right	on	off	stop	go
go	0.0	93.3	70.0	90.0	100.0	90.0	40.0	66.7	90.0	83.3
stop	86.7	0.0	83.3	96.7	93.3	86.7	46.7	46.7	83.3	100.0
off	93.3	96.7	0.0	86.7	100.0	96.7	80.0	100.0	100.0	100.0
on	76.7	96.7	70.0	0.0	83.3	76.7	70.0	53.3	93.3	90.0
right	96.7	100.0	93.3	86.7	0.0	100.0	70.0	70.0	96.7	86.7
left	70.0	76.7	90.0	80.0	100.0	0.0	86.7	63.3	76.7	93.3
down	36.7	56.7	83.3	86.7	66.7	86.7	0.0	76.7	76.7	73.3
up	73.3	83.3	96.7	96.7	93.3	93.3	100.0	0.0	100.0	90.0
no	90.0	93.3	100.0	93.3	100.0	93.3	80.0	90.0	0.0	100.0
yes	90.0	96.7	86.7	96.7	86.7	90.0	66.7	56.7	96.7	0.0

Confusion matrix showing the efficacy of our targeted adversarial attacks on speech recognition model

- 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.

Attack Labeled as Source	Attack Labeled as Target	Attack Labeled as Other
89%	0.6%	9.4%

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

Code

Audio Samples

Bibliography

1. Speech commands dataset. https://research.googleblog.com/2017/08/launching-speech-commands-dataset.html
2. N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou. Hidden voice commands. In USENIX Security Symposium, pages 513–530, 2016.
3. I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.
4. N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In Security and Privacy (EuroS&P), 2016 IEEE European Symposium on, pages 372–387. IEEE, 2016.