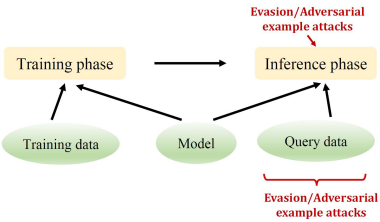
Zero-Query Adversarial Attack on Black-box Automatic Speech Recognition Systems

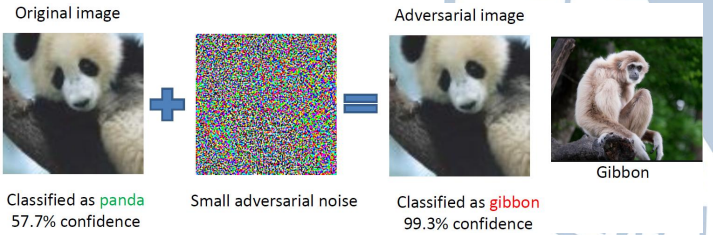
PhilFan
www.philfan.cn

1st Nov 2024

Zero-Query Adversarial Attack on Black-box Automatic Speech Recognition Systems

Adversarial Attack: change behavior to avoid detection





Zero-Query Adversarial Attack on Black-box Automatic Speech Recognition Systems

White-box attack

– Attackers have full knowledge about the ML model.

Black-box attack

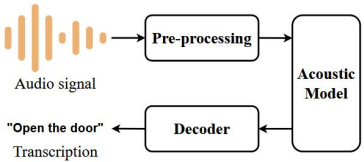
– Attackers don't have access to the ML model parameters, gradients,architecture

– Know about used ML algorithm

– **Zero-Query:Don't get query samples and query results**

Zero-Query Adversarial Attack on Black-box Automatic Speech Recognition Systems

spoken language ⟹ text



Target System

- **Online Speech Recognition**
- **Commercial IVC**
- **Open-source ASRs**

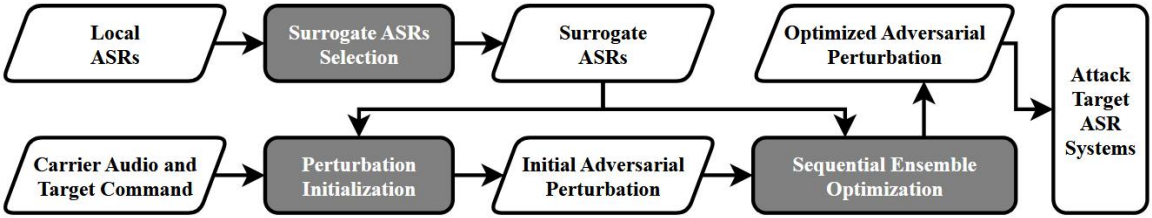
Jasper QuartzNet, ContextNet (M/L) , Citrinet (M/L) .....

Difficulties in Audio Adversarial Attacks

The target model is a black box, with no access to its structure, parameters, or training data. When using the target model for image classification, typically only a single label is provided without accuracy information. Additionally, API query limits impose cost constraints and potential detection by platform anomaly programs.

Speech systems must handle temporal information changes, which is more complex than image classification. Audio sampling rates are usually high (e.g., 16kHz, implying 16,000 samples per second), whereas images have only hundreds or thousands of pixels (e.g., 28×28 for MNIST and 32×32 for CIFAR-10).

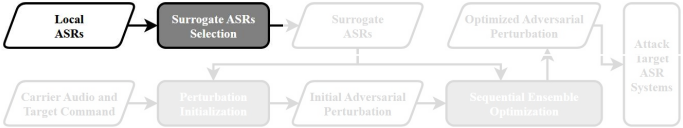
Workflow & Loss Function



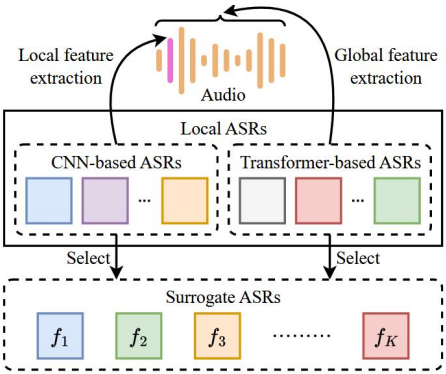
$$\max_{\delta} \mathop{\mathbb{P}}_{f \in \mathcal{F}} (f(x+\delta) = t)$$

$$\min_{\delta} \mathcal{L}_{all}(x, \delta, t, \mathbb{F}) \quad \text{s.t. } Dis(x, x') < \epsilon,$$

Surrogate ASRs Selection

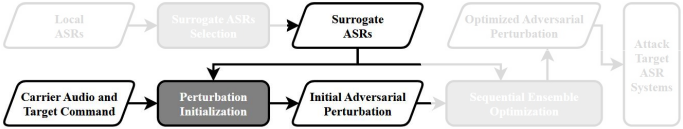


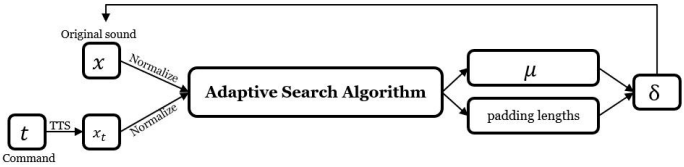
	Local features	Global information
CNNs	✓	✗
Transformers	✗	✓

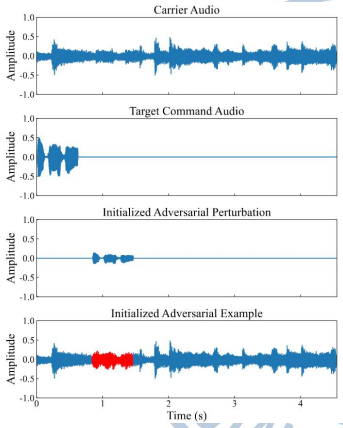


Incorporate both CNN-based and Transformer-based ASRs

Pertubation Initialization



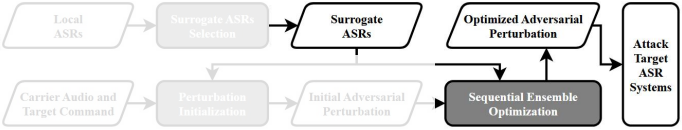


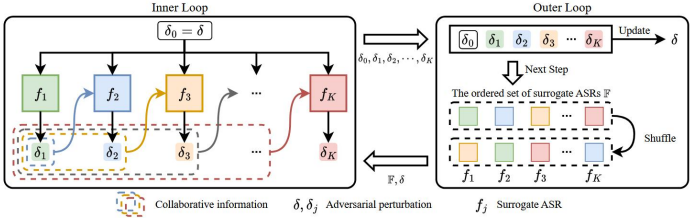


According to the algorithm, I draw this map to help us to understand

the adaptive search algorithm initializes δ by pushing adversarial example toward the decision boundary of all surrogate ASRs, thereby circumventing the time consuming and uncertain initial search process.

Sequential Ensemble Optimization





Inner Loop

$\delta_j = \delta_0 - \alpha \cdot \frac{1}{j} \sum_{\delta' \in \Delta_j} \nabla_{\delta'} \mathcal{L}(x, \delta' + \sigma, t, f_j)$ Updating one-by-one

$clip_\epsilon(\delta, x) = \max(\min(\delta, \epsilon \cdot |x|), -\epsilon \cdot |x|).$

restrict the updated δj within a limited range

Outer Loop

- Update δ
- Shuffle the set of surrogate ASRs

considers both current & preceding surrogate ASRs.

Experiment Design

Hardware


Geforce 3080Ti ×8

Intel Xeon Gold 5117 ×2


iPhone 13


Amazon Echo Dot

Baseline

Over-the-line: Carlini,Occam,KENKU.
Over-the-air: NI-Occam and KENKU.

SRoA & SNR

Data

Online Speech Recognition

Method	Azure	Tencent	Alibaba	OpenAI	Average	SNR (dB) ↑	Query ↓
Carlini *et al.* [10]	0/10	0/10	0/10	0/10	0/10	/	0
Occam [69]	10/10	10/10	10/10	10/10	10/10	12.54	30000
KENKU [62]	10/10	8/10	0/10	9/10	6.75/10	12.72	>0
ZQ-Attack	**10/10**	**10/10**	**10/10**	**10/10**	**10/10**	**21.91**	**0**

Commercial IVC

Method	Siri	alexa	Average	SNR (dB) ↑
NI-Occam [69]	4/10	5/10	4.5/10	8.38
KENKU [62]	7/10	9/10	8/10	12.72
ZQ-Attack	**10/10**	**10/10**	**10/10**	**15.77**

Open-source ASRs

Target ASR	SRoA	SNR (dB)	Target ASR	SRoA	SNR (dB)
Jasper	10/10	13.59	Conformer-CTC (XL)	10/10	23.59
QuartzNet	10/10	12.96	Conformer-Transducer (M)	10/10	25.34
Citrinet (M)	10/10	14.67	Conformer-Transducer (L)	10/10	20.63
Citrinet (L)	10/10	15.89	Conformer-Transducer (XL)	10/10	21.08

Data

Command	Azure		Tencent		Alibaba		OpenAI	
	Attack	SNR (dB)	Attack	SNR (dB)	Attack	SNR (dB)	Attack	SNR (dB)
ask me a question	✓	23.51	✓	28.66	✓	26.31	✓	28.73
clear notification	✓	26.52	✓	21.28	✓	20.05	✓	26.52
close the shades	✓	26.54	✓	26.24	✓	26.54	✓	26.86
find a hotel	✓	25.78	✓	24.63	✓	20.26	✓	28.21
good morning	✓	19.65	✓	18.96	✓	27.90	✓	25.76
I have a secret to tell you	✓	27.21	✓	27.21	✓	27.21	✓	27.21
I need help	✓	27.23	✓	26.64	✓	20.17	✓	28.54
open the box	✓	18.48	✓	26.34	✓	18.48	✓	27.54
read a book	✓	20.77	✓	25.81	✓	17.27	✓	27.71
record a video	✓	21.38	✓	20.51	✓	21.38	✓	21.38
reset password	✓	22.70	✓	21.69	✓	19.91	✓	22.70
show me my message	✓	27.12	✓	27.54	✓	23.98	✓	27.54
show me the money	✓	27.94	✓	27.96	✓	25.75	✓	27.96
start recording	✓	27.03	✓	20.42	✓	19.87	✓	27.19
tell me a story	✓	27.89	✓	27.89	✓	24.34	✓	27.94
turn off the fan	✓	19.98	✓	15.25	✓	19.98	✓	19.98
turn on the TV	✓	25.55	✓	17.71	✓	17.18	✓	25.55
watch TV	✓	26.32	✓	26.32	✓	19.20	✓	26.32
what time is it	✓	26.22	✓	25.39	✓	25.39	✓	28.20
where is my car	✓	27.24	✓	23.28	✓	25.76	✓	27.24
Average	20/20	24.75	20/20	23.99	20/20	22.35	20/20	26.45

Command	NI-Occam		KENKU		ZQ-Attack	
	Siri	Alexa	Siri	Alexa	Siri	Alexa
call my wife	✓	✓	✓	✓	✓	✓
make it warmer	✗	✗	✗	✓	✓	✓
navigate to my home	✓	✓	✓	✓	✓	✓
open the door	✗	✗	✓	✓	✓	✓
open the website	✗	✓	✓	✓	✓	✓
play music	✓	✓	✓	✓	✓	✓
send a text	✗	✗	✗	✗	✓	✓
take a picture	✗	✗	✗	✓	✓	✓
turn off the light	✓	✓	✓	✓	✓	✓
turn on airplane mode	✗	✗	✓	✓	✓	✓
Average	4/10	5/10	7/10	9/10	10/10	10/10

Evaluation on a large command set

Thanks