CommanderSong: A Systematic Approach For Practical Adversarial Voice Recognition

Xuejing Yuan[1,2], Yuxuan Chen[3], Yue Zhao[1,2], Yunhui Long[4], Xiaokang Liu[1,2], Kai Chen[1,2], Shengzhi Zhang[3, 5], Heqing Huang, XiaoFeng Wang[6], and Carl A. Gunter[4]

[1]SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, China
[3]Department of Computer Sciences, Florida Institute of Technology, USA
[4]Department of Computer Science, University of Illinois at Urbana-Champaign, USA
[5]Department of Computer Science, Metropolitan College, Boston University, USA
[6]School of Informatics and Computing, Indiana University Bloomington, USA

目　录

1 研究背景——智能语音

1 研究背景——智能语音


iFLY


herman/kardon
amazon
amazon


Capital One
PLATINUM
5160 9656 7890 1234
01/17
R THOMAS
MasterCard
THE LOUVRE MUSEUM
B M W

1 研究背景——AI安全

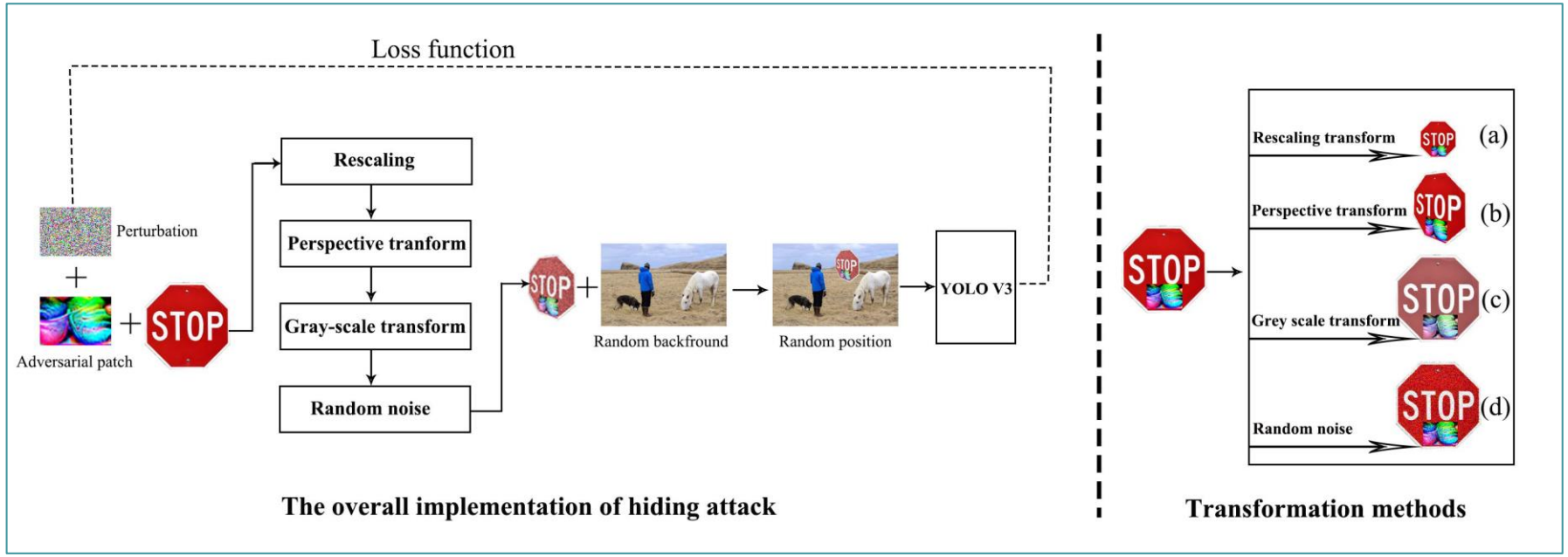
- 传统攻击AI方法
  - 恶意应用
  - 病毒传播
  - 无线劫持
  - 固件重写

1 研究背景——AI安全

- 新型攻击AI方法
  - 数据污染
  - 窃取模型

1 研究背景——AI安全

- 新型攻击AI方法
  - 数据污染
  - 窃取模型
  - 对抗样本



[1] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. Computer Vision and Pattern Recognition, 2018.

[2] Zhao Y, Zhu H, Shen Q, et al. Practical Adversarial Attack Against Object Detector[J]. arXiv preprint arXiv:1812.10217, 2018.

目　录

2 相关工作——智能语音系统攻击

[1] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In USENIX Security Symposium, pages 513–530, 2016.

2 相关工作——智能语音系统攻击

[1] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In USENIX Security Symposium, pages 513–530, 2016.

[2] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pages 103–117. ACM, 2017.

2 相关工作——智能语音系统攻击

[1] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In USENIX Security Symposium, pages 513–530, 2016.

[2] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pages 103–117. ACM, 2017.

[3] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. Deep Learning and Security Workshop, 2018.

目　录

3 CommanderSong攻击方案

3 CommanderSong攻击方案

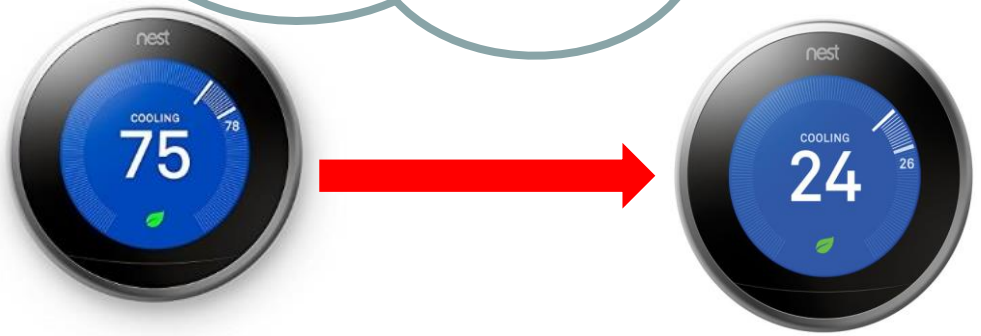
I think the command in the sound wants me to set the temprature to 24 degree.

----IVC device

3 CommanderSong攻击方案

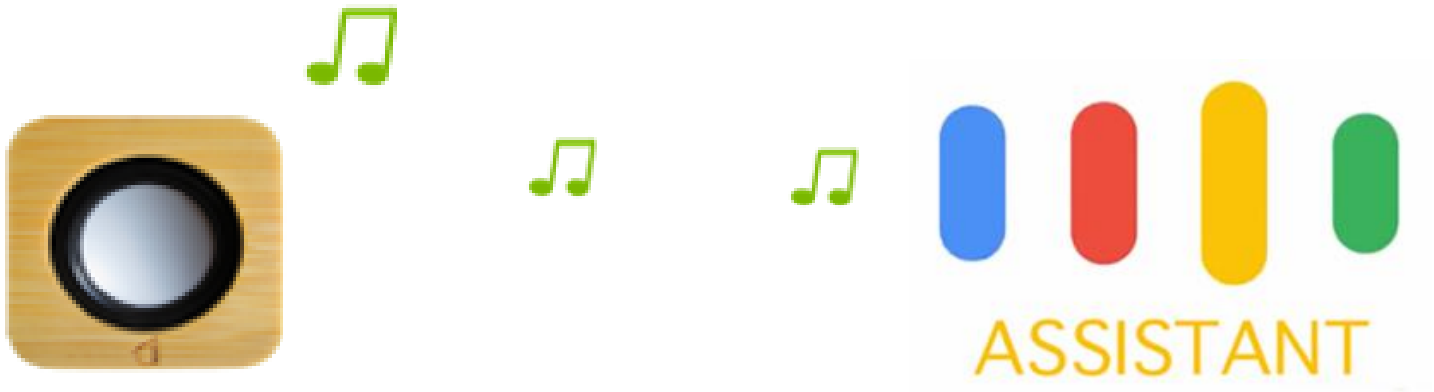
I think the command in the sound wants me to set the temprature to 24 degree.

----IVC device



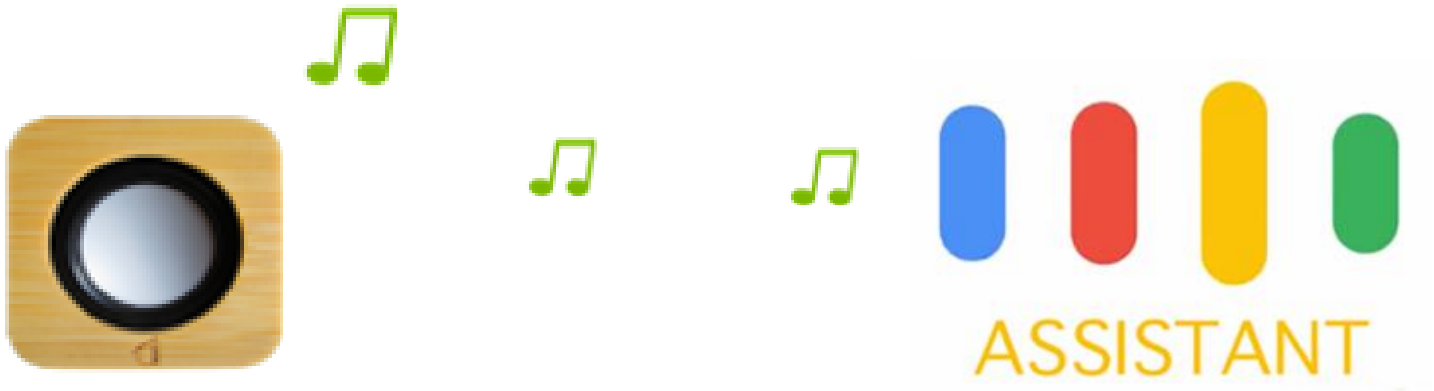
3 CommanderSong攻击方案

- WTA (WAV-To-API) 攻击
- WAA (WAV-Air-API) 攻击



3 CommanderSong攻击方案

- WTA (WAV-To-API) 攻击
- WAA (WAV-Air-API) 攻击



开源语音识别平台Kaldi

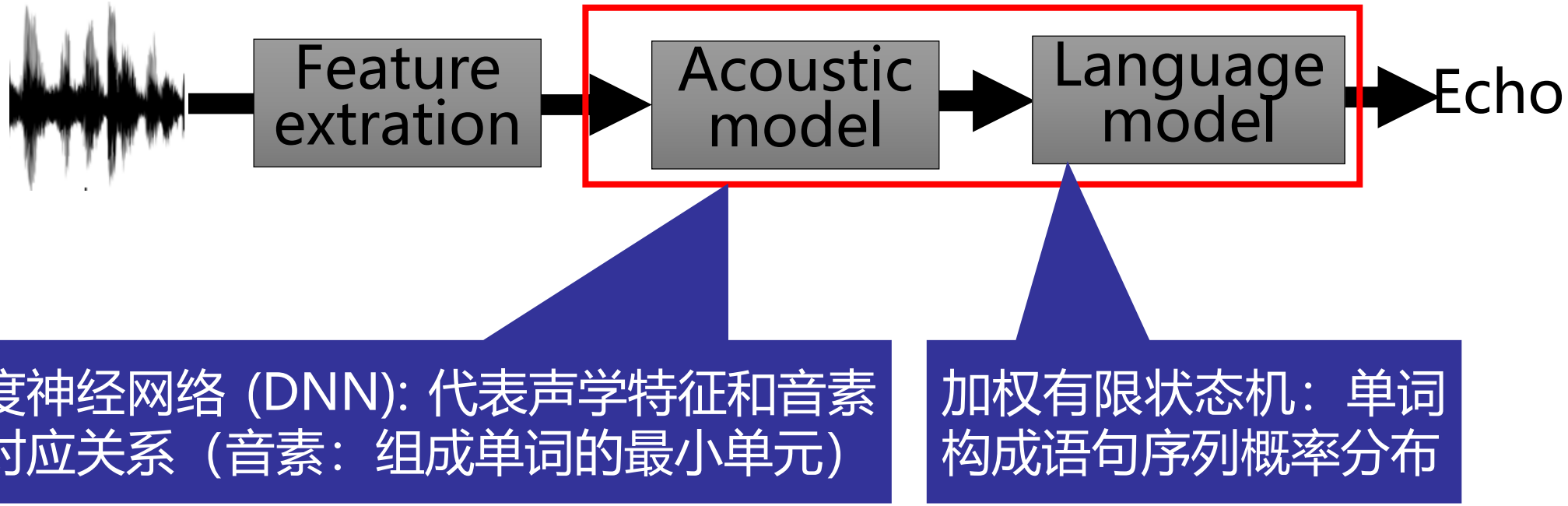
目　录

4 CommanderSong攻击原理——语音识别原理

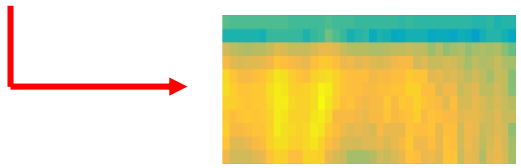
4 CommanderSong攻击原理——语音识别原理

深度神经网络 (DNN): 代表声学特征和音素的对应关系（音素：组成单词的最小单元）

4 CommanderSong攻击原理——语音识别原理



4 CommanderSong攻击原理——语音识别原理

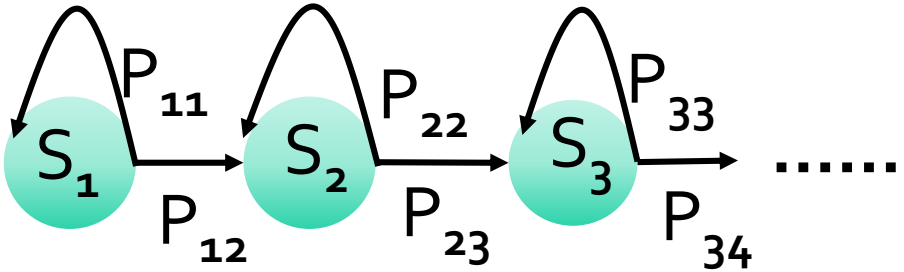


①

$O_1O_2O_3O_4\cdots$

观察状态

②



(音素的HMM状态转换关系)

③

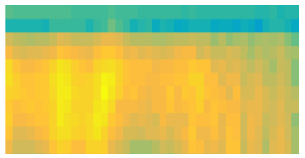
4 CommanderSong攻击原理——语音识别原理

Feature extration

Acoustic model

Language model

Echo



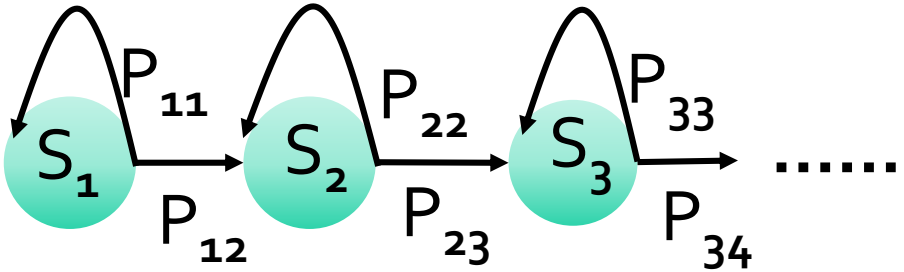
①

pdf-id：音素和声学特征概率分布对应关系 (DNN输出矩阵的列序号)

$O_1O_2O_3O_4\cdots$

观察状态

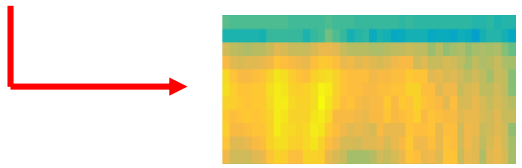
②



(音素的HMM状态转换关系)

③

4 CommanderSong攻击原理——语音识别原理



①

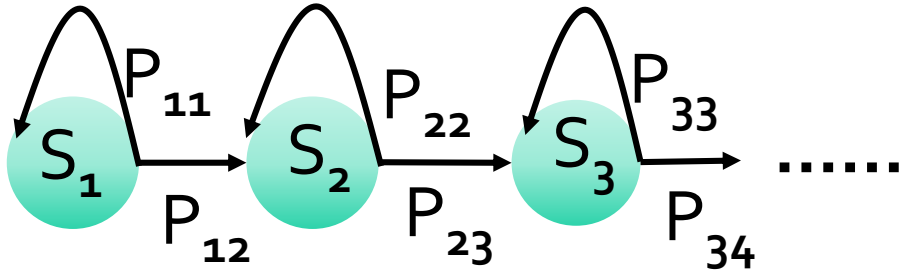
pdf-id：音素和声学特征概率分布对应关系 (DNN输出矩阵的列序号)

$O_1O_2O_3O_4\cdots$

观察状态

②

transition-id：音素HMM的状态转换关系



(音素的HMM状态转换关系)

③

4 CommanderSong攻击原理——语音识别原理

eh_B
15985_16190_16189_16189_16189_16189_16189_16189_16189_16189
k_I
31123_31380_31379_31379_31379_31379_31379_31379_31379_31379
ow_E
39643_39898_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897

单词“Echo”解码得到的Transition-ids序列

音素、HMM状态、pdf-id以及transition-id关系对照表

Phoneme	HMM state	Pdf-id	Transition-id	Transition
eh_B	0	6383	15985	0→1
			15986	0→2
eh_B	1	5760	16189	self-loop
			16190	1→2

4 CommanderSong攻击原理——语音识别原理

eh_B
15985_16190_16189_16189_16189_16189_16189_16189_16189_16189

k_I
31123_31380_31379_31379_31379_31379_31379_31379_31379_31379

ow_E
39643_39898_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897

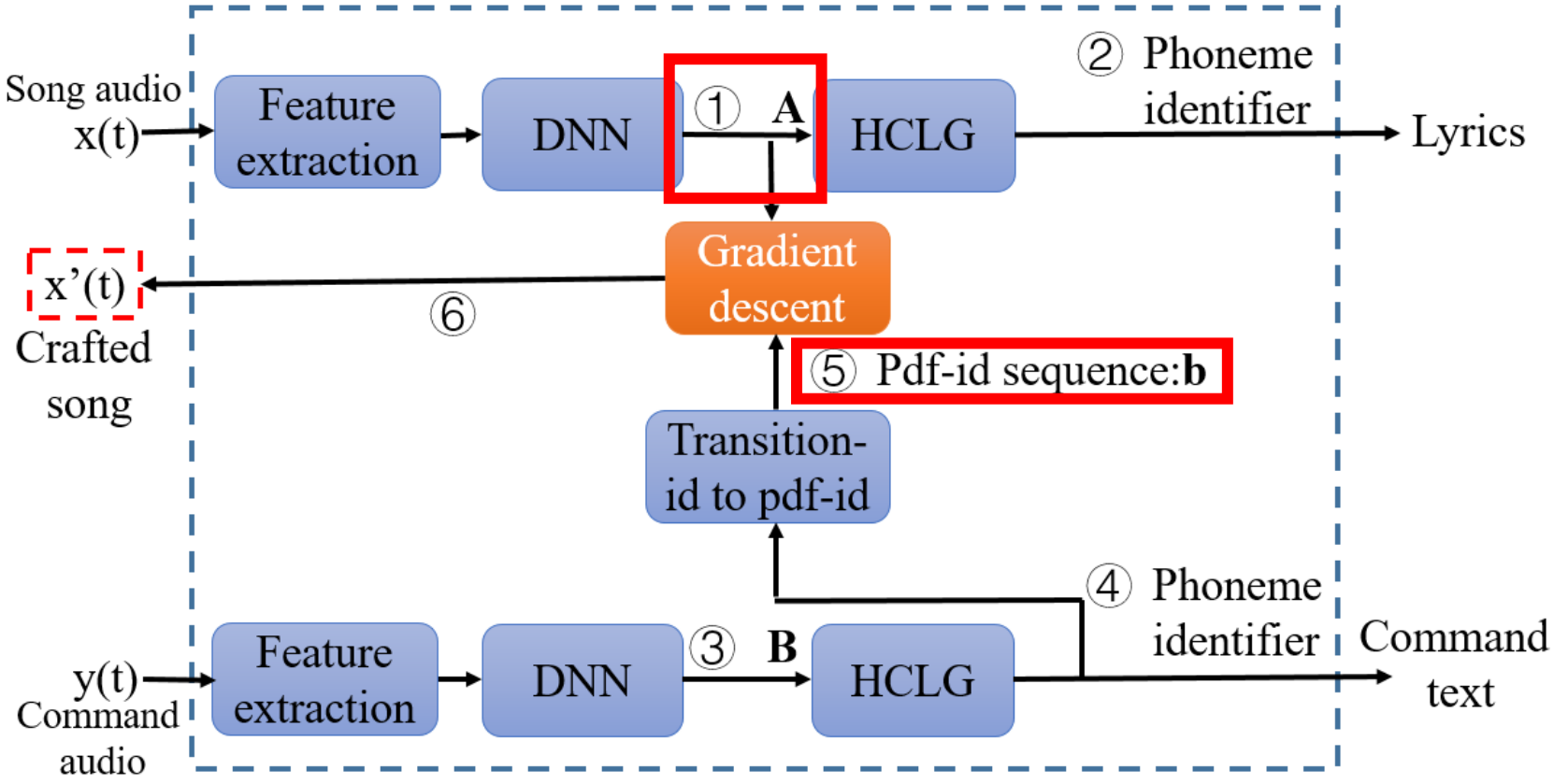
单词“Echo”解码得到的Transition-ids序列

音素、HMM状态、pdf-id以及transition-id关系对照表

Phoneme	HMM state	Pdf-id	Transition-id	Transition
eh_B	0	6383	15985	0→1
			15986	0→2
eh_B	1	5760	16189	self-loop
			16190	1→2

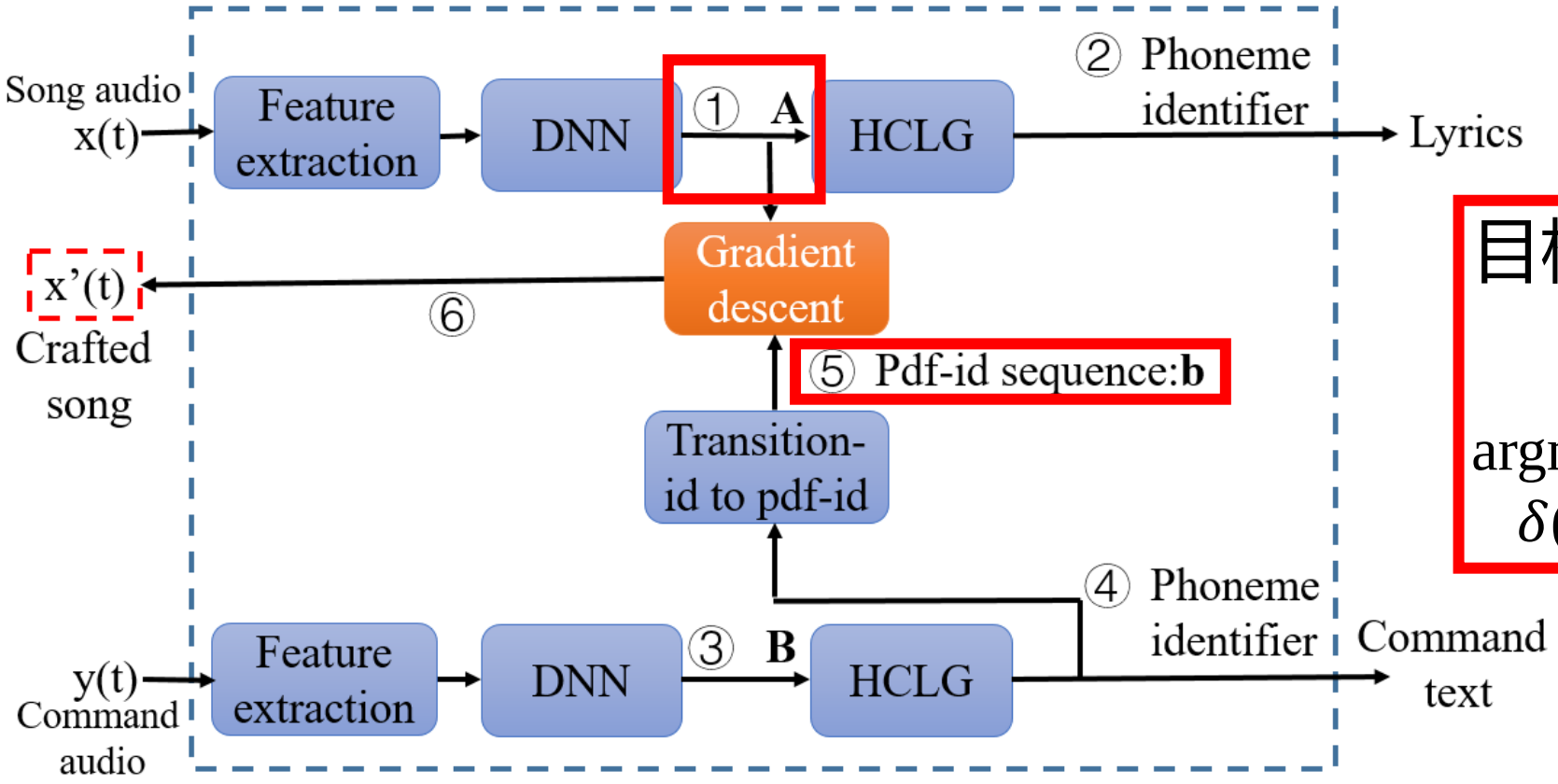
pdf-ids 序列： 6383, 5760,5760, 5760, 5760, 5760, 5760, 5760, 5760 .....

4 CommanderSong攻击原理——WTA攻击



Pdf-id 匹配算法

4 CommanderSong攻击原理——WTA攻击



Pdf-id 匹配算法

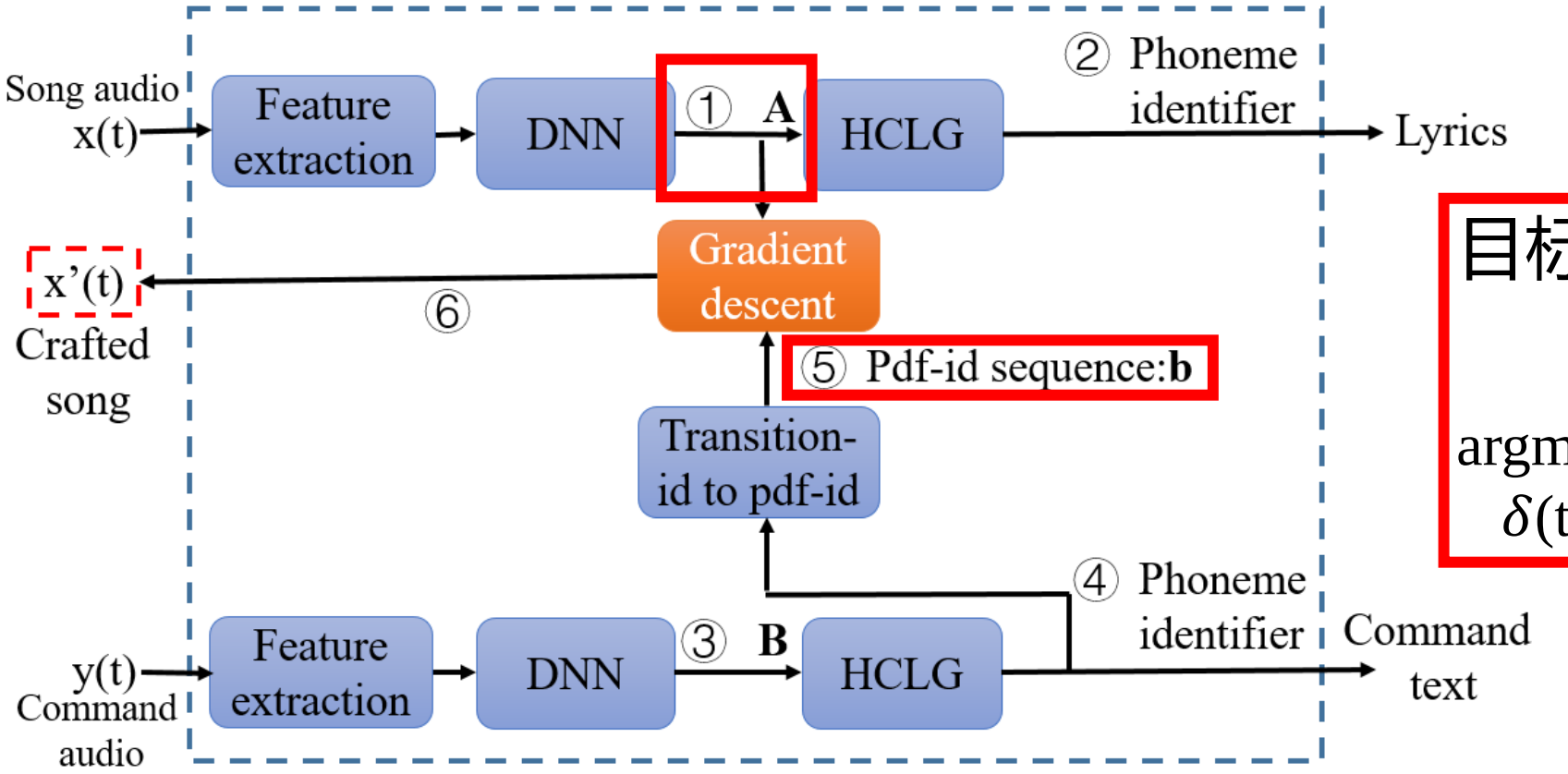
目标函数:

$$m_i = \arg\max a_{i,j}\ ,$$

$$g(x(t))=\mathbf{m}.$$

$$\underset{\delta(t)}{\operatorname{argmin}} \|g(x(t)+\delta(t))-\mathbf{b}\|_1.$$

4 CommanderSong攻击原理——WTA攻击



Pdf-id 匹配算法

目标函数:

$$m_i = \arg\max a_{i,j},$$

$$g(x(t))=\mathbf{m}.$$

$$\underset{\delta(t)}{\operatorname{argmin}} \|g(x(t)+\delta(t))-\mathbf{b}\|_1.$$

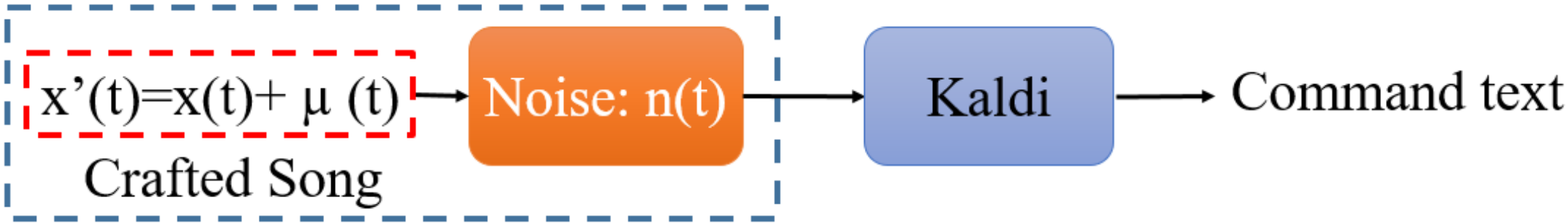
WTA 攻击成功!

4 CommanderSong攻击原理——WAA攻击

- 噪声模型（模拟背景噪声和设备电子噪声）
- 随机噪声模型

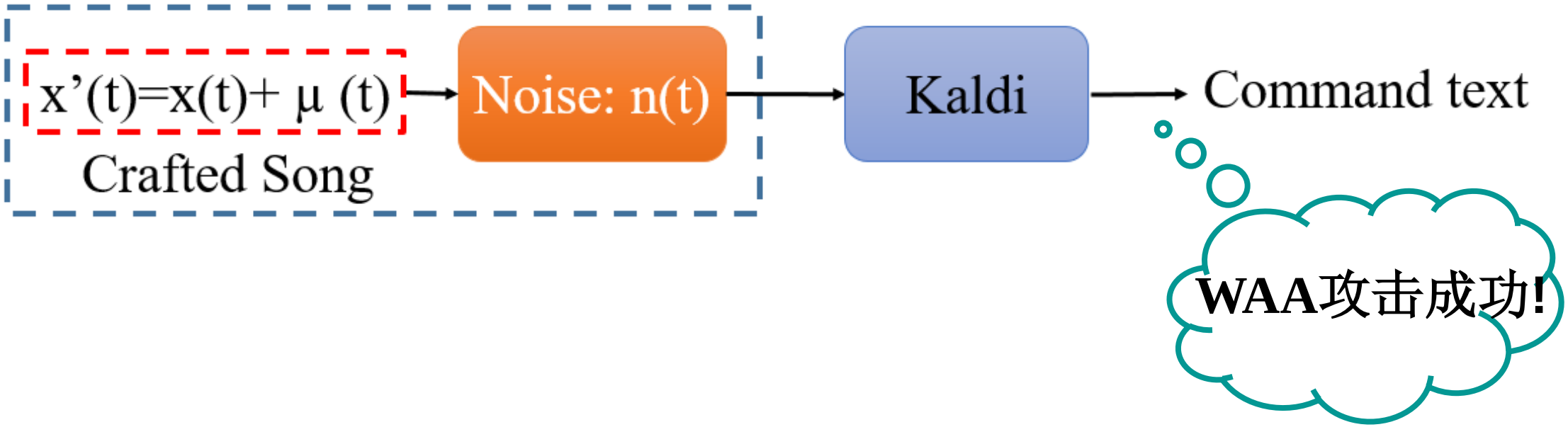
4 CommanderSong攻击原理——WAA攻击

- 噪声模型（模拟背景噪声和设备电子噪声）
- 随机噪声模型



4 CommanderSong攻击原理——WAA攻击

- 噪声模型（模拟背景噪声和设备电子噪声）
- 随机噪声模型



目　录

5 实验评估

WTA 攻击结果

Command	Success rate (%)
Okay google restart phone now.	100
Okay google flashlight on.	100
Okay google read mail.	100
Okay google clear notification.	100
Okay google airplane mode on.	100
Okay google turn on wireless hot spot.	100
Okay google read last sms from boss.	100
Echo open the front door.	100
Echo turn off the light.	100

5 实验评估

WAA 攻击结果

Command	Speaker	Success rate (%)
Echo ask capital one to make a credit card payment.	JBL speaker	90
	ASUS Laptop	82
	SENMATE Broadcast	72
Okay google call one one zero one one nine one two zero.	JBL speaker	96
	ASUS Laptop	60
	SENMATE Broadcast	70

5 实验评估

人类对WTA攻击样本的理解

Music classification	Listened (%)	Abnormal (%)	Recognize command (%)
Soft music	13	15	0
Rock	33	28	0
Popular	32	26	0
Rap	41	23	0

5 实验评估

人类对WAA攻击样本的理解

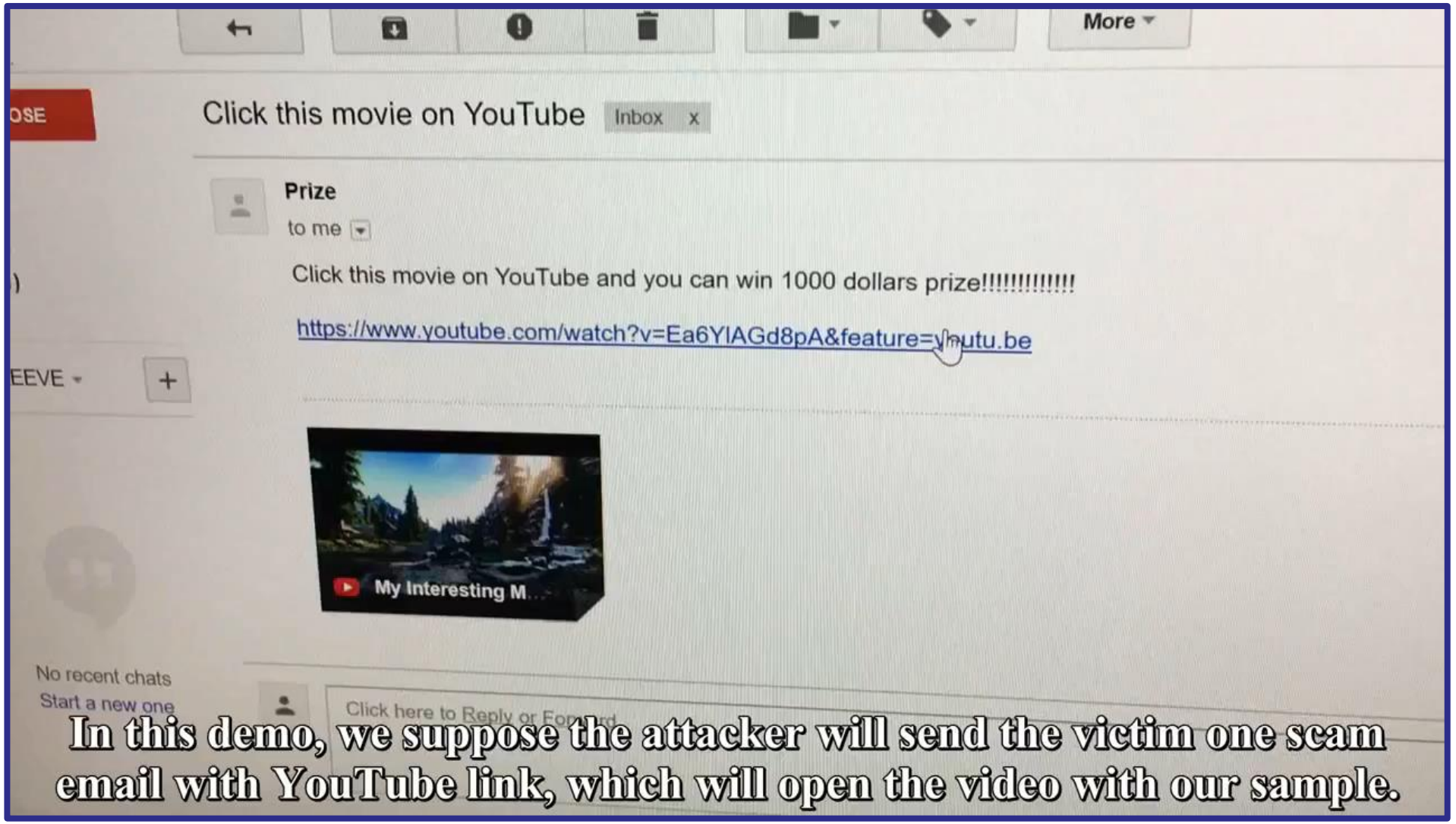
Song name	Listened (%)	Abnormal (%)	Noise-speaker (%)	Noise-song (%)
Did You Need It	15	67	42	1
Outlaw of Love	11	63	36	2
The Saltwater Room	27	67	39	3
Sleepwalker	13	67	41	0
Under neath	13	68	45	3
Feeling Good	38	59	36	4
Average	19.5	65.2	40	2.2

5 实验评估

CommanderSong攻击科大讯飞

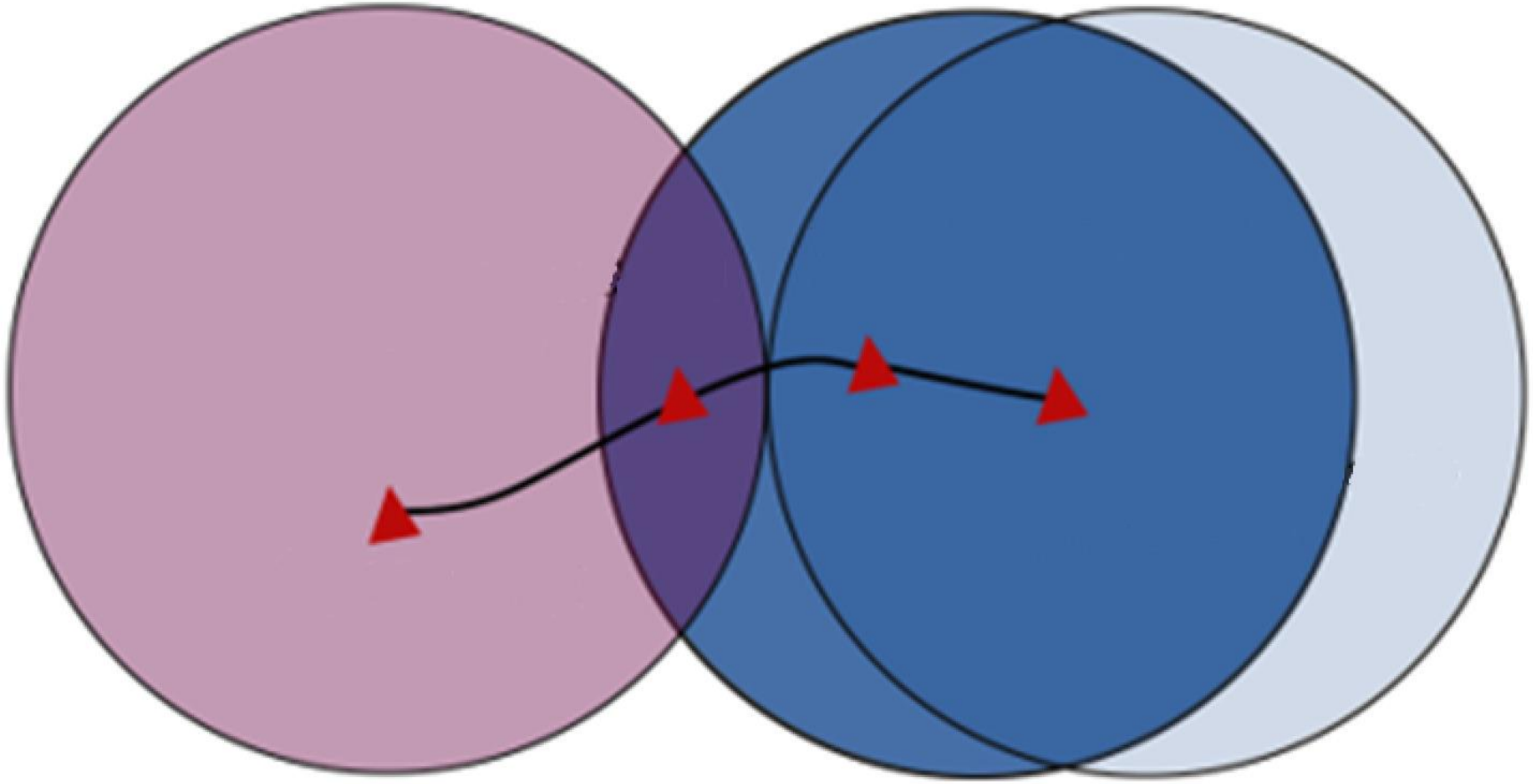
Command	iFLYREC (%)	iFLYTEK Input (%)
Airplane mode on.	66	0
Open the door.	100	100
Good night.	100	100

5 实验评估


More
Click this movie on YouTube
Inbox x
Prize
to me
Click this movie on YouTube and you can win 1000 dollars prize!!!!!!!!!!!!!
https://www.youtube.com/watch?v=Ea6YlAGd8pA&feature=youtu.be
My Interesting M...
No recent chats
Start a new one
Click here to Reply or Forward
In this demo, we suppose the attacker will send the victim one scam email with YouTube link, which will open the video with our sample.

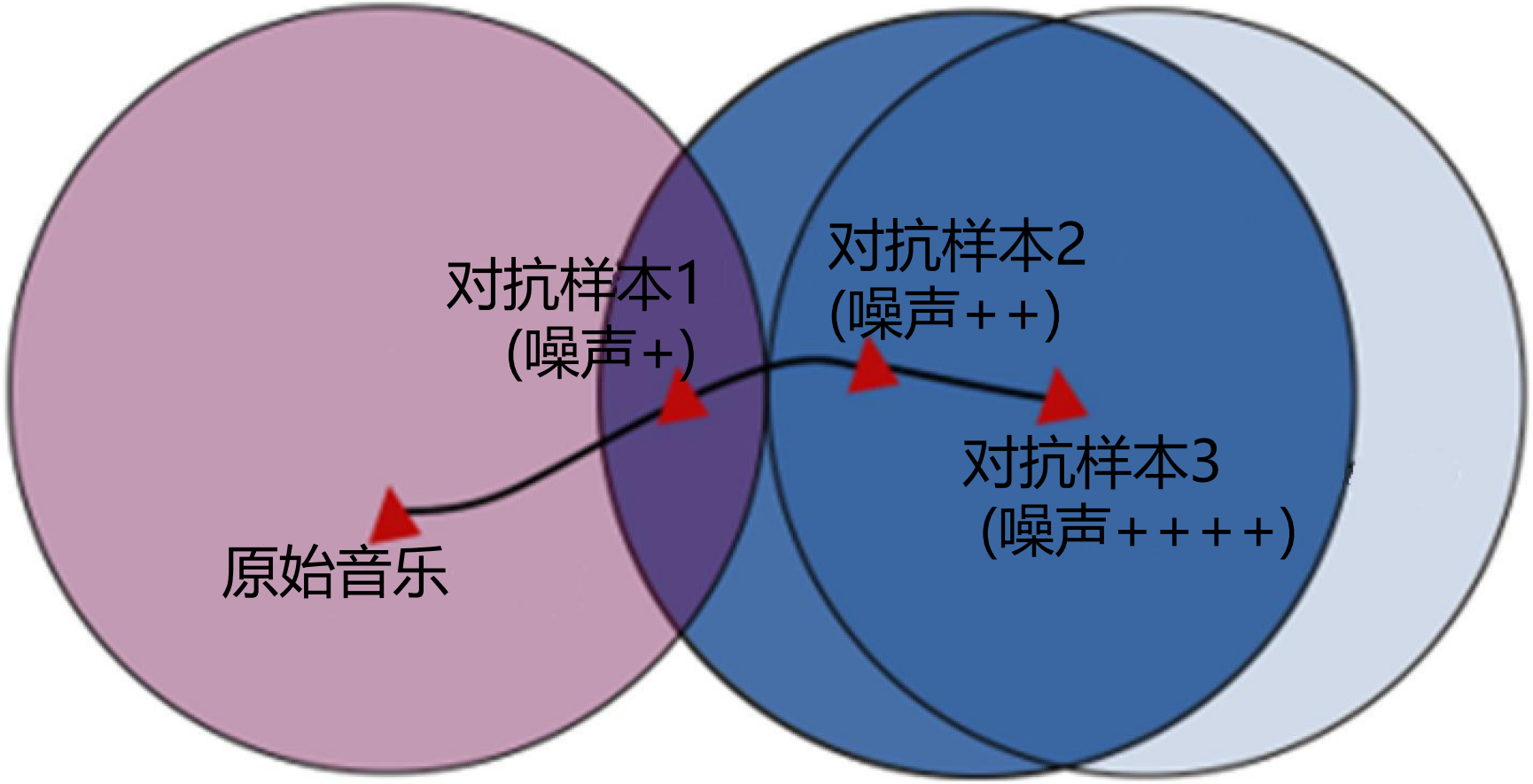
5 实验评估

- 人类识别成歌曲
- Kaldi 识别成指令
- 人类识别成指令



5 实验评估

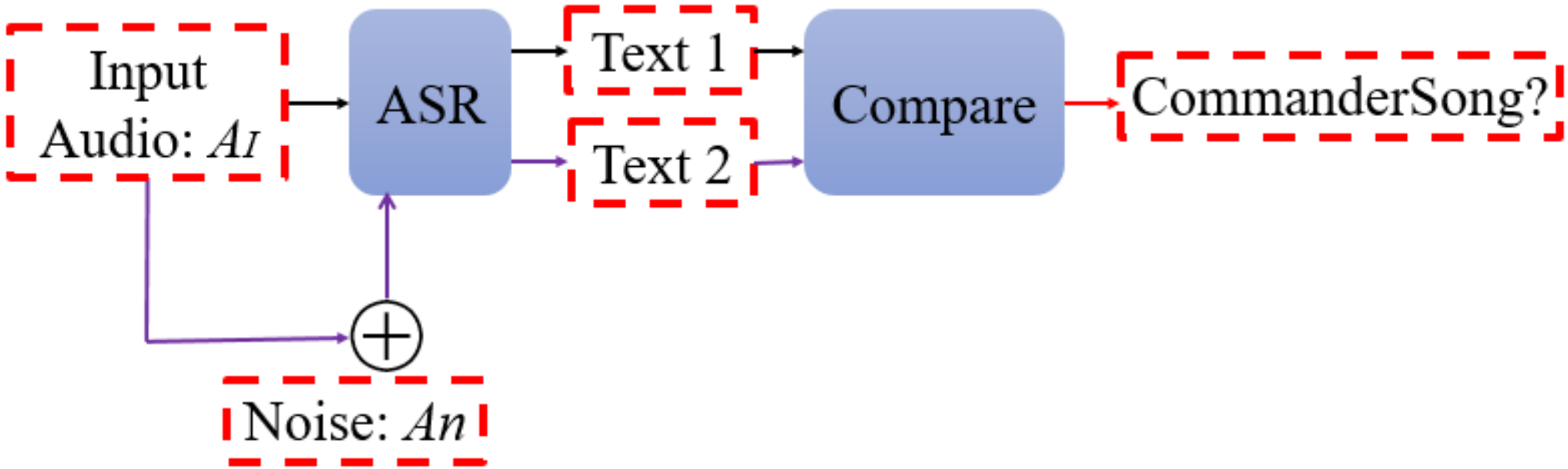
- 人类识别成歌曲
- Kaldi 识别成指令
- 人类识别成指令



目　录

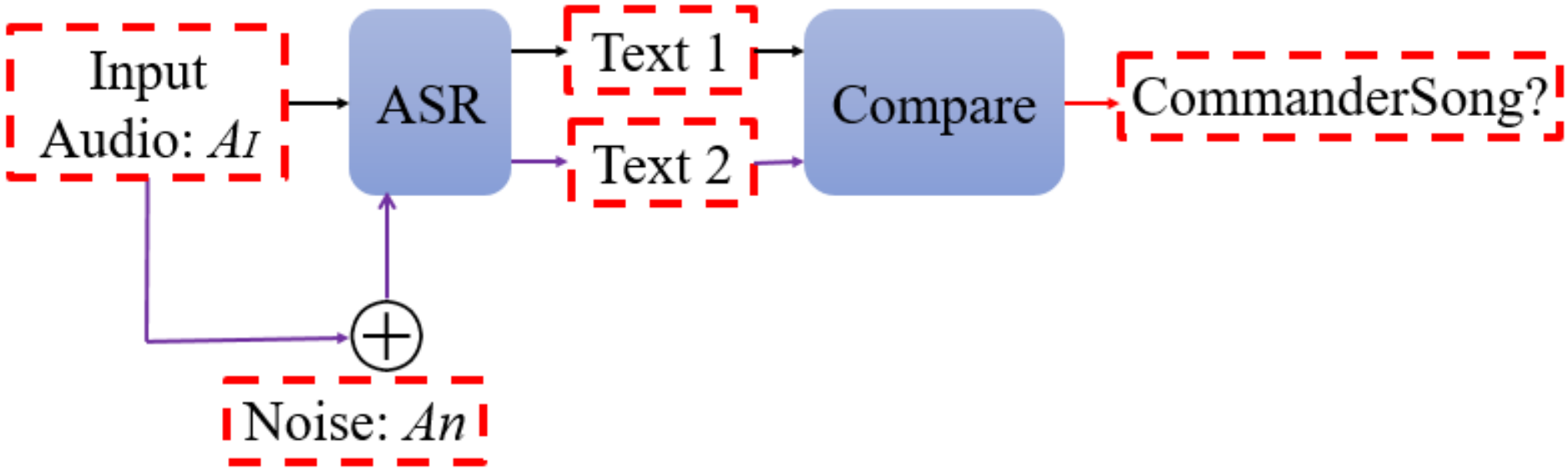
6 攻击防御

- 音频干扰防御



6 攻击防御

- 音频干扰防御



- 音频压缩防御

总结

- 成果：
  - 实际对抗攻击语音识别系统
  - 攻击商业化平台（科大讯飞）
  - 通过网络或者无线信号传播
  - 人类难以察觉

总结

- 成果：
  - 实际对抗攻击语音识别系统
  - 攻击商业化平台（科大讯飞）
  - 通过网络或者无线信号传播
  - 人类难以察觉

- 思路：逆向语音识别算法，从神经网络的误判找突破口。
- 难点：分析神经网络误判可能区域，克服物理环境影响。


谢 谢 ！