Attack on Speaker Recognition

Todo List

Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems

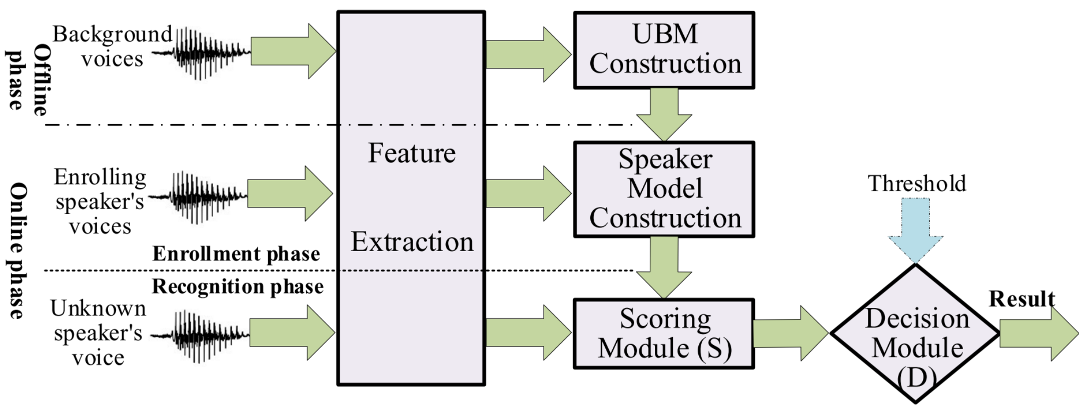
Contribution

1. 实现了对说话人识别的对抗攻击, 将说话人识别中的判别阈值很好地加入到对抗样本的生成过程中;
2. 针对黑盒, 实现了有目标/无目标地攻击攻击;
3. 添加的扰动非常的少, 实现的效果可观;
4. 进行了大量的实验;
5. 这个攻击的一个缺点是需要依赖 API 输出相应的标签概率;

Notes

1. **黑盒**的, **物理/API**的, **有/无目标**的说话人识别对抗攻击;
2. 说话人识别模型:

   (1) 经典的 UBM-GMM 模型

   (2) 说话人识别处理的任务:

   - Open-set Identification (OSI): 识别为哪一个说话人或返回空;

- Close-set Identification (CSI): 识别为其中一个说话人 (不会返回空);
- Speaker Verification (SV): 验证是否是目标说话人;

(3) 是否依赖文本: 从后面的实验来看, 依赖文本的语音识别系统可能具有更好的安全性;

- 依赖文本;
- 不依赖文本;

(4) 模型结构:

- ivector-PLDA;
- GMM-UBM;
- xvector-PLDA;

3. 威胁场景:
   - 攻击黑盒模型;
   - 黑盒模型需要**输出识别的结果和得分**, 如果没有得分的话, 就使用迁移攻击 (如在 Microsoft Azure 上);
   - 👍 作者总共考虑 16 中可能的攻击组合:

$$\left\{\begin{array}{c}\begin{pmatrix}\text{targeted}\\ \text{untargeted}\end{pmatrix}\times\begin{pmatrix}\text{intra-gender}\\ \text{inter-gender}\end{pmatrix}\times \text{API}\times\begin{pmatrix}\text{OSI}\\ \text{CSI}\\ \text{SV}\end{pmatrix}\times \text{D.\&S.}\\ +\\ \text{targeted}\times\begin{pmatrix}\text{OSI}\\ \text{CSI}\\ \text{SV}\end{pmatrix}\times \text{API}\times \text{decision-only}\\ +\\ \text{targeted}\times\begin{pmatrix}\text{OSI}\\ \text{CSI}\\ \text{SV}\end{pmatrix}\times \text{over-the-air}\times \text{D.\&S.}\\ +\\ \text{targeted}\times \text{OSI}\times \text{over-the-air}\times \text{decision-only}\end{array}\right\}$$

4. 👍 算法:

(1) 迭代算法的选择: NES 算法是梯度估计算法 (**梯度估计算法的特点是需要知道目标标签的概率**) 中最佳的, **PSO** 算法是遗传算法中最佳的, 这里作者选用的是 **NES** 算法;

(2) 形式化问题:

$$\begin{array}{l}\operatorname{argmin}_{\delta} f(x+\delta)\\ \text{such that } \|x+\delta, x\|_{\infty}< \epsilon \text{ and } x+\delta\in[-1,1]^n\end{array}$$

在一定扰动范围内, 是的目标 loss 函数最小化;

(3) Attack on OSI:

- Targeted Attack:

$$f(x)=\max\left\{(\max\{\theta, \max_{i\in G\setminus\{t\}}[S(x)]_i\}-[S(x)]_t), -\kappa\right\}$$

最大化目标概率, 是的目标概率超过阈值 θ , 添加一个系数 k 增强样本的鲁棒性, k 越大越鲁棒.

- Untargeted Attack: (文章的公式可能有点小错误)

$$f(x)=\max\{(\theta-\max_{i\in G\setminus\{t\}}[S(x)]_i), -k\}$$

这一块作者**并没有考虑 reject 也是无目标攻击的一种**, 故会有上面这个式子. 如果转换为 **平常我们遇到的无目标攻击 (考虑 reject)**, 公式形式如下:

$$f(x)=\max\{[S(x)]_t-\theta, -k\}$$

即我们让标签小于 θ 就完成了无目标攻击, 但如果这样的话, 便无法和下面的 θ **估计算法** 相结合, 因为我们这里需要对 θ 向下估值, 而非向上估值.

- θ 估计算法:

Algorithm 1 Threshold Estimation Algorithm

Input: The target OSI system with scoring S and decision D modules
An arbitrary voice x such that $D(x) = \texttt{reject}$

Output: Estimated threshold $\acute{\theta}$

1: $\acute{\theta} \leftarrow \max_{i \in G}[S(x)]_i$; ▷ *initial threshold*
2: $\Delta \leftarrow |\frac{\acute{\theta}}{10}|$; ▷ *the search step*
3: $\acute{x} \leftarrow x$;
4: **while** True **do**
5: $\quad \acute{\theta} \leftarrow \acute{\theta} + \Delta$;
6: $\quad f' \leftarrow \lambda x. \max\{\acute{\theta} - \max_{i \in G}[S(x)]_i, -\kappa\}$; ▷ *loss function*
7: $\quad$ **while** True **do**
8: $\quad\quad \acute{x} \leftarrow \texttt{clip}_{x,\epsilon}\{\acute{x} - \eta \cdot \texttt{sign}(\nabla_x f'(\acute{x}))\}$; ▷ *craft sample using* f'
9: $\quad\quad$ **if** $D(\acute{x}) \neq \texttt{reject}$ **then**; ▷ $\max_{i \in G}[S(\acute{x})]_i \geq \theta$
10: $\quad\quad\quad$ **return** $\max_{i \in G}[S(\acute{x})]_i$;
11: $\quad\quad$ **if** $\max_{i \in G}[S(\acute{x})]_i \geq \acute{\theta}$ **then break**;

大致的思想是, 先初始化一个较小的估计值 $\acute{\theta}$, 如果迭代生成对抗样本超过了这个估计值, 但却未输出目标说明人标签时, 增大估计值继续生成对抗样本; (伪代码第 6 行的 λx 挺奇怪的, 没太理解)

- 梯度估计 - NES 算法:

$$\frac{1}{m \times \sigma} \sum_{j=1}^{m} f(\acute{x}_{i-1}^{j}) \times u_j$$

其中, $u_j = -u_{m+1-j}$, σ 是高斯分布的方差;

- 梯度更新 - BIM 算法:

$$\acute{x}_i = \texttt{clip}_{x,\epsilon}\{\acute{x}_{i-1} - \eta \cdot \texttt{sign}(\nabla_x f(\acute{x}_{i-1}))\}$$

- 参数选择: $m = 50$, $\delta = 1e-3$, $\eta \in [1e-3, 1e-6]$, $max\ iteration = 1000$;

(4) Attack on CSI: 和 OSI 不同指出是, CSI 一定会输出一个标签, 因此不需要考虑 θ 的问题

- Targeted Attack:

$$f(x) = \max\left\{(\max_{i \in G \setminus \{t\}}[S(x)]_i - [S(x)]_t), -\kappa\right\}$$

- Untargeted Attack:

$$f(x) = \max\{([S(x)]_m - \max_{i \in G \setminus \{m\}}[S(x)]_i), -\kappa\}$$

(5) Attack on SV: SV 是一个单分类的识别系统, 如果为目标说话人则返回 True, 否则返回 False, 因此这种攻击下没有 Targeted / Untargeted 之分.

$$f(x) = \max\{\theta - S(x), -\kappa\}$$

这里将非目标说话人的语音转化为目标说话人的标签;

5. 👍 Evaluation on Effectiveness and Efficiency:

(1) 数据集:

Datasets	#Speaker	Details
Train-1 Set	7,273	Part of **VoxCeleb1** [69] and whole **VoxCeleb2** [70] used for training ivector and GMM
Train-2 Set	2,411	Part of **LibriSpeech** [71] used for training system C in transferability
Test Speaker Set	5	5 speakers from **LibriSpeech** 3 female and 2 male, 5 voices per speaker, voices range from 3 to 4 seconds
Imposter Speaker Set	4	Another 4 speakers from **LibriSpeech** 2 female and 2 male, 5 voices per speaker, voices range from 2 to 14 seconds

(2) 评价指标:

Metric	Description
Attack success rate (ASR)	Proportion of adversarial voices that are recognized as the target speaker
Untargeted success rate (UTR) for CSI	Proportion of adversarial samples that are not recognized as the source speaker
Untargeted success rate (UTR) for OSI	Proportion of adversarial samples that are not rejected by the target system

(3) 本地训练的黑盒模型: 设置阈值参数 $\theta_{ivector} = 1.45$, $\theta_{GMM} = 0.091$ 以保证 FAR 在 10% 左右；

Task	Metrics	ivector	GMM
CSI	Accuracy	99.6%	99.3%
SV	FRR	1.0%	5.0%
	FAR	11.0%	10.4%
OSI	FRR	1.0%	4.2%
	FAR	7.9%	11.2%
	OSIER	0.2%	2.8%

- FRR : False Rejection Rate;
- FAR : False Acceptance Rate;
- $OSIER$: Open Set Identification Error Rate is the rate of voices that can not be correctly classified;

(3) 修改量的大小: 实验中选择 $\epsilon = 0.002$;

ϵ	ivector				GMM			
	#Iter	Time (s)	SNR (dB)	ASR (%)	#Iter	Time (s)	SNR (dB)	ASR (%)
0.05	18	422	12.0	100	18	91	16.7	100
0.01	23	549	16.2	100	16	81	19.1	100
0.005	44	1099	21.8	100	19	102	22.3	100
0.004	56	1423	23.8	100	21	104	24.0	100
0.003	76	2059	26.3	100	27	124	26.1	100
0.002	124	2845	30.2	99	40	218	29.3	99
0.001	276	6738	36.4	41	106	551	35.7	87

(4) 攻击结果: **相比之下, ivector的对抗样本更难生成, 最少的一个样本需要迭代 25 轮(即 query 1250 次)**;

Task	System								System (Intra-gender attack)								System (Inter-gender attack)							
	ivector				GMM				ivector				GMM				ivector				GMM			
	#Iter	Time (s)	SNR (dB)	ASR (%)	#Iter	Time (s)	SNR (dB)	ASR (%)	#Iter	Time (s)	SNR (dB)	ASR (%)	#Iter	Tine (s)	SNR (dB)	ASR (%)	#Iter	Time (s)	SNR (dB)	ASR (%)	#Iter	Time (s)	SNR (dB)	ASR (%)
CSI	124	2845	30.2	99.0	40	218	29.3	99.0	92	2115	29.3	100.0	25	126	28.8	100.0	146	3340	30.8	98.0	50	278	29.62	98.0
SV	84	2014	31.6	99.0	39	241	31.4	99.0	31	751	31.7	98.0	30	185	31.7	100.0	135	3252	31.6	100.0	48	298	31.2	98.0
OSI	86	2277	31.5	99.0	38	226	31.4	99.0	32	833	31.3	98.0	31	178	31.5	100.0	140	3692	31.6	100.0	45	274	31.2	98.0

(5) 过程中得到的阈值估计:

ivector			GMM		
θ	$\acute{\theta}$	Time (s)	θ	$\acute{\theta}$	Time (s)
1.45	**1.47**	**628**	**0.091**	**0.0936**	**157**
1.57	1.60	671	0.094	0.0957	260
1.62	1.64	686	0.106	0.1072	269
1.73	1.75	750	0.113	0.1141	289
1.84	1.87	804	0.119	0.1193	314

(6) 攻击 Talentedsoft 平台: 成功攻击;

6. Evaluation on Transferability:

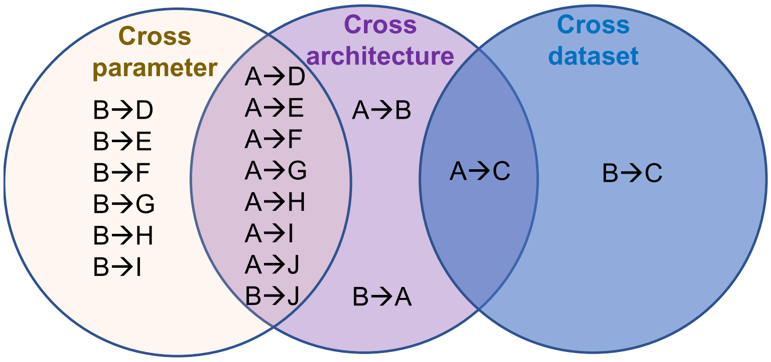
(1) 目标模型结构: A, B, J 为前面实验用到的模型, 这边针对 ivector 和 GMM 增加了 `C~I` 模型;

System ID	A	B	C	D	E	F	G	H	I	J
Architecture	**GMM**	**ivector**	ivector	ivector	ivector	ivector	ivector	ivector	ivector	**xvector**
Training set	Train-1 Set	Train-1 Set	**Train-2 Set**	Train-1 Set	Train-1 Set	Train-1 Set	Train-1 Set	Train-1 Set	Train-1 Set	Train-1 Set
Feature	MFCC	MFCC	MFCC	**PLP**	MFCC	MFCC	MFCC	MFCC	**PLP**	MFCC
DF	24×3	24×3	24×3	24×3	**13×3**	24×3	24×3	24×3	**13×3**	**30**
FL/FS (ms)	25/10	25/10	25/10	25/10	25/10	**50/10**	25/10	25/10	**50/10**	25/10
♯GC	2048	2048	2048	2048	2048	2048	**1024**	2048	**1024**	–
DV	–	400	400	400	400	400	400	**600**	**600**	**512**

(2) 目标模型训练结果:

Task \ System		C	D	E	F	G	H	I	J
CSI	**Accuracy**	99.8%	99.4%	99.2%	99.8%	99.6%	99.8%	99.2%	99.2%
SV	**FAR**	10.0%	9.8%	9.4%	10.0%	11.2%	9.8%	10.4%	10.2%
	FRR	1.2%	0.6%	1.6%	1.2%	0.8%	1.0%	2.2%	0.8%
OSI	**FAR**	9.1%	8.8%	10.9%	9.2%	8.5%	8.1%	11.0%	7.7%
	FRR	1.4%	0.6%	1.6%	1.4%	1.2%	0.8%	2.2%	0.8%
	OSIER	0.0%	0.2%	0.2%	0.0%	0.2%	0.0%	0.4%	0.2%

(3) Transferability 的种类: 包括 跨平台, 跨模型种类 和 跨数据集;



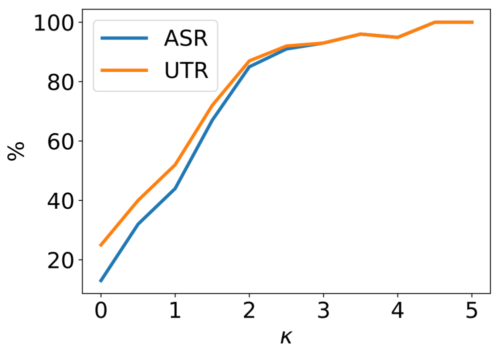
(4) 为了提高 Transfer 能力, 作者对参数的设置如下:

- 修改量: $\epsilon = 0.05$, 可以看到在迁移攻击中需要更大的修改量;
- CSI Task: $k_{GMM} = 0.2$, $k_{ivector} = 10$;
- SV Task: $k_{GMM} = 3$, $k_{ivector} = 4$;
- OSI Task: $k_{GMM} = 3$, $k_{ivector} = 5$;

(5) 实验结果:

S \ T	A		B		C		D		E		F		G		H		I		J	
	ASR	**UTR**	**ASR**	**UTR**	**ASR**	**UTR**	**ASR**	**UTR**	**ATR**	**UTR**	**ASR**	**UTR**	**ASR**	**UTR**	**ASR**	**UTR**	**ASR**	**UTR**	**ASR**	**UTR**
A	—	—	62.0	64.0	48.0	48.0	55.2	56.9	68.0	68.0	64.0	64.0	52.0	54.0	68.0	68.0	38.0	40.0	34.0	42.0
B	5.0	5.0	—	—	67.5	67.5	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	72.5	75.0	40.0	41.7

(6) 讨论 k 的影响: k 越大, transferability 的能力越好;



(7) 攻击 Microsoft Azure 平台: 由于 Azure 上不输出相应的概率, 因此使用 transfer 攻击.

- Text-Independent OSI-Azure:
- ☆ Text-Dependent SV-Azure: 只实现了 10% 的成功率, 其他的都因为添加的噪声过多而出现 "**Error, too noisy**";

> 文本相关的, 并且判断语音中的噪声, 这样的语音认证对于攻击更加鲁棒, 能否对这个系统进行攻击?

7. Evaluation on Over-the-Air

(1) 实验环境:

	System	Loudspeaker	Microphone	Distance	Acoustic Environment
Different Systems	**GMM OSI/CSI/SV ivector OSI/CSI/SV Azure OSI**	JBL clip3 portable speaker	IPhone 6 Plus (iOS)	1 meter (65 dB)	relatively quiet
Different Devices	ivector OSI	**DELL laptop JBL clip3 portable speaker Shinco brocast equipment**	**IPhone 6 Plus (iOS) OPPO (Android)**	1 meter (65 dB)	relatively quiet
Different Distances	ivector OSI	JBL clip3 portable speaker	IPhone 6 Plus (iOS)	**0.25 meter (70 dB) 0.5 meter (68 dB) 1 meter (65 dB) 2 meters (62 dB) 4 meters (60 dB) 8 meters (55 dB)**	relatively quiet
Different Acoustic Environments	ivector OSI	JBL clip3 portable speaker	IPhone 6 Plus (iOS)	1 meter (65 dB)	**white noise (45/50/60/65/75 dB) bus noise (60 dB) restaurant noise (60 dB) music noise (60 dB) absolute music noise (60 dB)**

(2) Result of Different Systems:

System		SNR (dB)	Result (%)	
			Normal voices	**Adversarial voices**
ivector	**CSI**	6.6	Accuracy: 100	ASR: 80, UTR: 80
	SV	9.8	FAR: 0, FRR: 0	ASR: 76
	OSI	7.8	FAR: 4, FRR: 0, OSIER: 0	ASR: 100, UTR: 100
GMM	**CSI**	6.1	Accuracy: 85	ASR: 90, UTR: 100
	SV	7.9	FAR: 0, FRR: 62	ASR: 100
	OSI	8.2	FAR: 0, FRR: 65, OSIER: 0	ASR: 100, UTR: 100
Azure	**OSI**	6.8	FAR: 5, FRR: 2, OSIER: 0	ASR: 70, UTR: 70

(3) Result of Different Devices:

L \ M	iPhone 6 Plus (iOS)					OPPO (Android)				
	Normal voices			Adv. voices		Normal voices			Adv. voices	
	FAR	**FRR**	**OSIER**	**ASR**	**UTR**	**FAR**	**FRR**	**OSIER**	**ASR**	**UTR**
DELL	10	0	0	100	100	13	6	0	78	80
JBL clip3	4	0	0	100	100	6	0	0	80	80
Shinco	8	5	0	89	91	14	0	0	75	75

(4) Result of Different Distance:

Distance (meter)		**0.25**	**0.5**	**1**	**2**	**4**	**8**
Normal Voices	**FAR**	4	3	4	6	0	0
	FRR	0	0	0	5	10	32
	OSIER	0	0	0	0	0	0
Adversarial Voices	**ASR**	100	100	100	70	40	10
	UTR	100	100	100	70	50	10

(5) Result of Different Acoustic Environment:

Environment		Quiet	White (45 dB)	White (50 dB)	White (60 dB)	White (65 dB)	White (75 dB)	Bus (60 dB)	Rest. (60 dB)	Music (60 dB)	Abs. Music (60 dB)
Normal voices	**FAR**	4	0	6	0	0	10	0	0	0	4
	FRR	0	5	12	30	**40**	**97**	25	20	10	10
	OSIER	0	0	0	0	0	0	0	0	10	0
Adv. voices	**ASR**	100	75	70	57	20	2	50	50	66	48
	UTR	100	75	70	60	20	2	50	50	67	48

Links

- 论文链接: [Chen, Guangke, et al. "Who is real bob? adversarial attacks on speaker recognition systems." *S&P* (2021).]
- 论文主页: https://sites.google.com/view/fakebob
- 论文代码: https://github.com/FAKEBOB-adversarial-attack/FAKEBOB