重放检测问题的深入分析

报告人：程星亮

目录

声纹识别系统（ASV）可能遭受攻击

人声模仿

- 模仿伪造韵律、口音等高阶特性
- 易欺骗人耳，难欺骗ASV系统
- 未见研究证明其具有显著威胁性

语音合成

- 基于规则 (共振峰)、数据 (单元选择)、参数(声带激励、声道调制)、端到端(频谱) 的合成技术
- 通过声码器 (GriffinLim等) 合成伪造语音进行攻击
- **需要一定技术，具有威胁性**

声音转换

录音重放

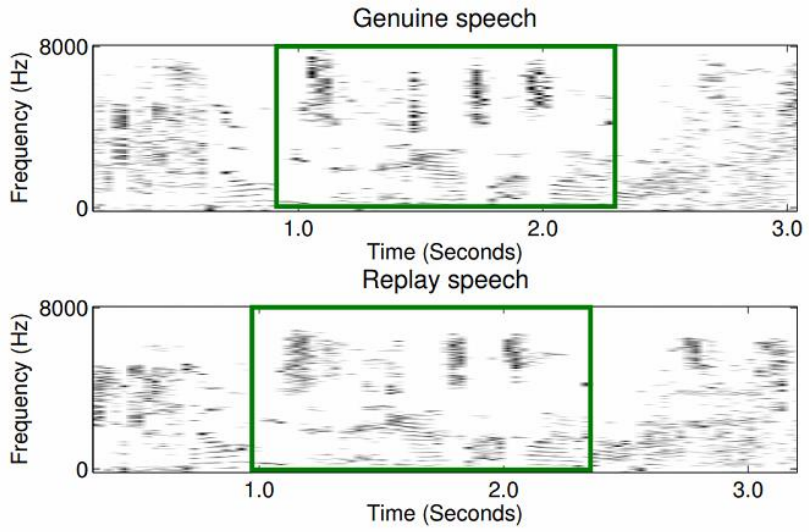
- 先用麦克风录制
- 然后用扬声器重放播放伪造语音进行攻击
- 最常见、易实施、威胁大

对抗样本

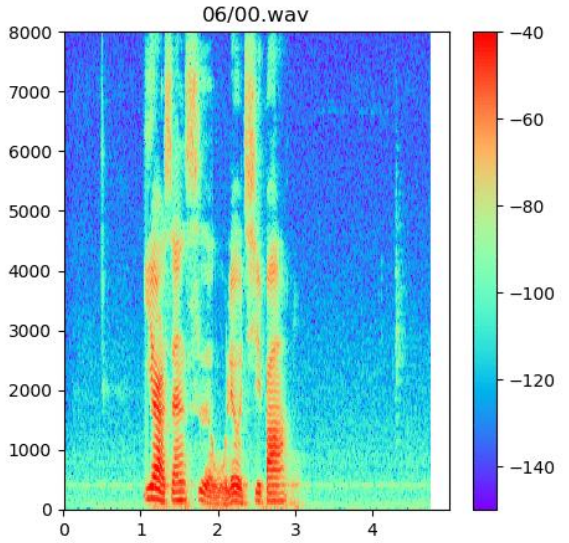
- 针对特定模型，用梯度方法微调语音
- 有动态语音对齐、黑盒泛化性等问题
- 实施难度高，一旦成功威胁巨大

应对重放攻击的方法

基于挑战-响应的方式

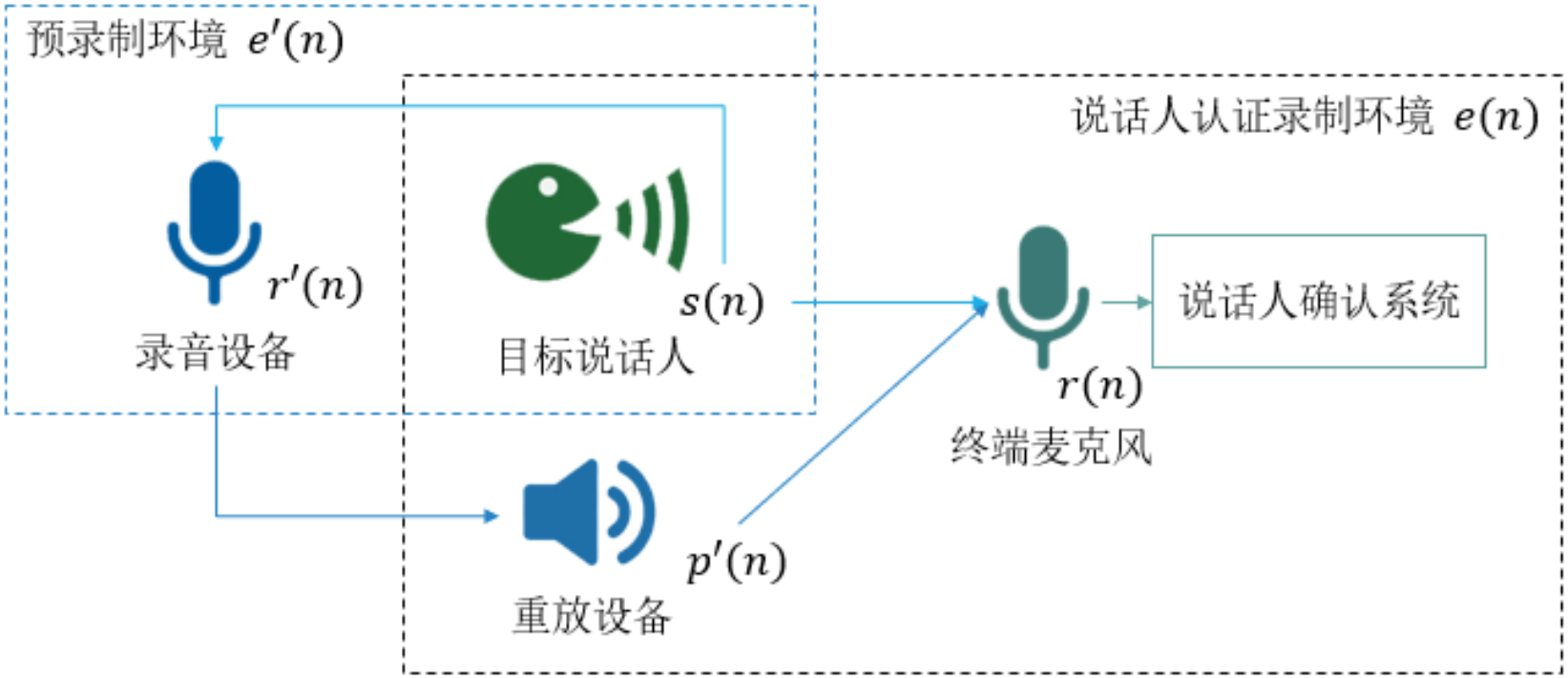


基于模板匹配的方式



重放失真检测

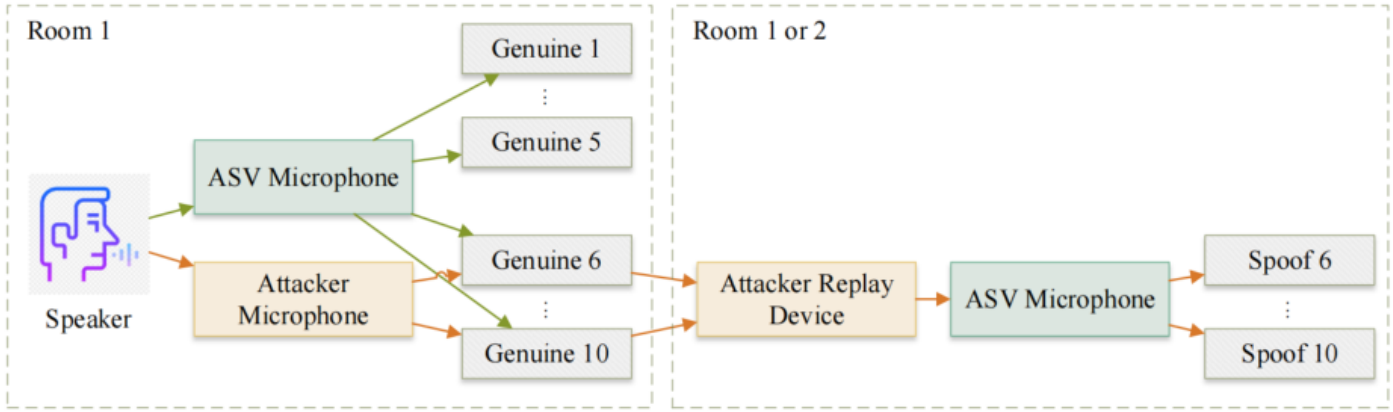
真实认证与重放攻击流程



期望：分析重放过程与不同因子之间的关系

要求： 1. 因子全交叉　　2. 除目标因子外其他扰动小

Parallel Replay Attack Digit (PRAD) 数据集



数据集录制流程图

(a) Replay session 1

(b) Replay session 2

(c) Replay session 3

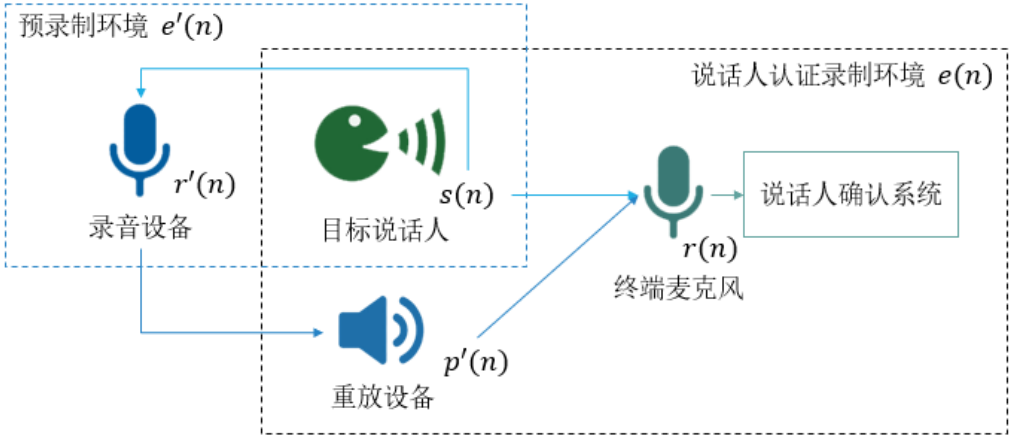
录制场景示例

数据集包含的设备列表

Device Type	ID	Model and Brand
Att Mic	R1	iPhone 6S
	R2	XiaoMi MI 5
	R3	Algo R6601
	R4	Algo R6688
	R5	Newsmy V29
	R6	Philips VTR7100
	R7	Rode NT1000
Att Replay	P1	Huawei P20 Pro
	P2	iPhone XS
	P3	Edifier MB200
	P4	Microlab B17
	P5	Philips SPA311
	P6	Brookstone Macaron 155579
	P7	Thinkpad E470 (built-in speaker)
	P8	Audioengine A5
ASV Mic	T1	Huawei Mate 20 X
	T2	Huawei Nova 3
	T3	Honor 6 H60-L01
	T4	JNN Q70
	T5	Amoi A1 USB
	T6	Sony UX560F
	T7	Philips VTR6900

传递函数分析

理论与假设



前提假设

- 真实语音和重放语音的接受终端相同（即传递函数只与重放过程相关）
- 重放过程被视为线性时不变系统

$$X_g(\omega) = H_g(\omega)S(\omega)$$

$$X_{re}(\omega) = H_{re}(\omega)S(\omega),$$

$$X_g(\omega) = H_m(\omega)H_{s\to m}(\omega)S(\omega)$$

$$X_{re}(\omega) = H_{ar\to m}(\omega)H_{am}(\omega)H_{am}(\omega)H_{s\to am}(\omega)S(\omega).$$

$H_m(\omega)$

终端　说话人到终端　现实中的语音

终端　重放到终端　重放设备　Att麦克风　现实中的语音

说话人到Att麦克风

传递函数：$H_{g\to re} = H_{am}H_{ar}\{H_{s\to am}H_{ar\to m}H_{s\to m}^{-1}\}$，　和说话人与终端设备无关

传递函数与因子的关系

$$H_{g\to re} = H_{am}H_{ar}\left\{H_{s\to am}H_{ar\to m}H_{s\to m}^{-1}\right\},$$

Factor	Impacted Term in Transfer Function
Speaker	$H_{s\to am}H_{s\to m}^{-1}$
Att Mic	$H_{s\to am}H_{am}$
Att Replay	$H_{ar}H_{ar\to m}$
ASV Mic	$H_{ar\to m}H_{s\to m}^{-1}$

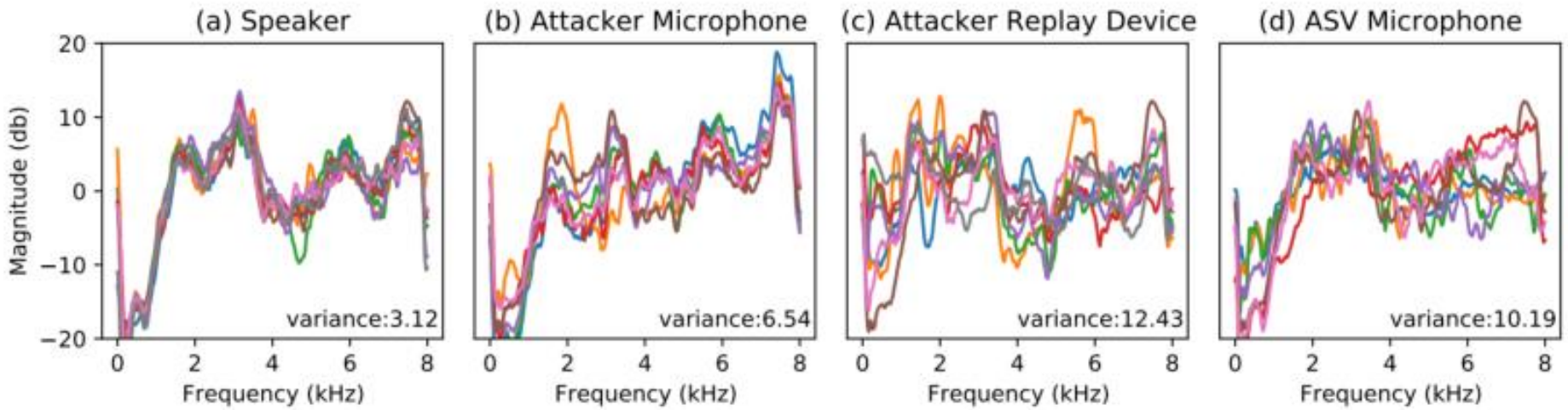
因素影响分析

$$H_{g\to re} = H_{am}H_{ar}\{H_{s\to am}H_{ar\to m}H_{s\to m}^{-1}\}$$

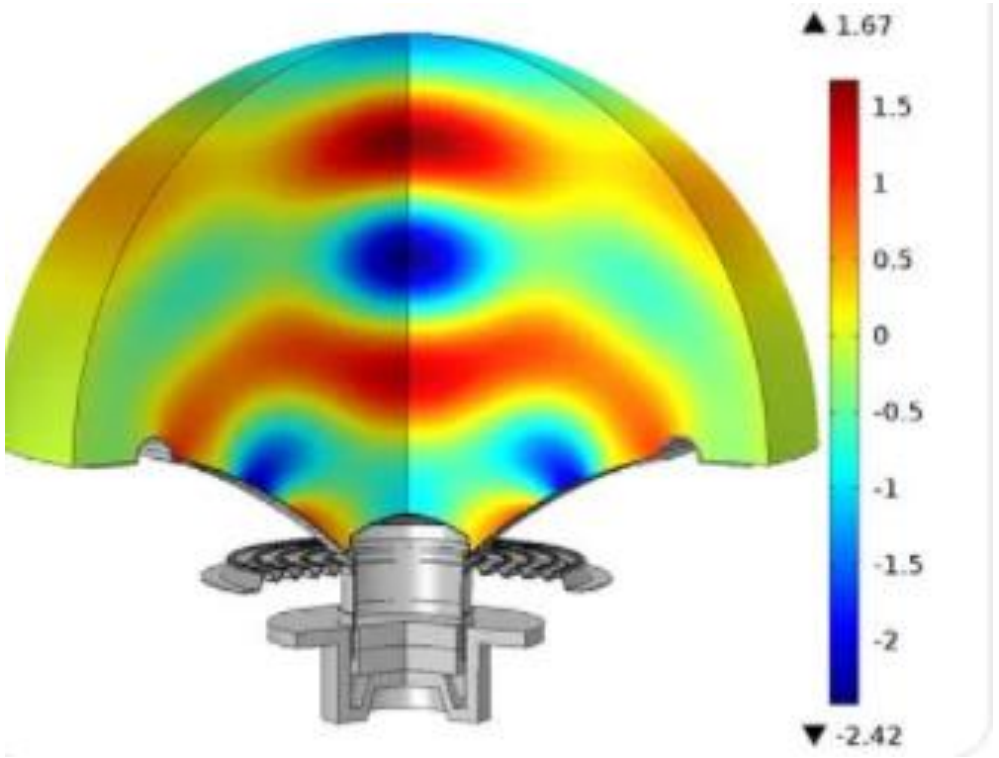
Factor	Impacted Term in Transfer Function
Speaker	$H_{s\to am}H_{s\to m}^{-1}$
Att Mic	$H_{s\to am}H_{am}$
Att Replay	$H_{ar}H_{ar\to m}$
ASV Mic	$H_{ar\to m}H_{s\to m}^{-1}$

$$\mathcal{F} = \{Speaker, AttMic, AttReplay, ASVMic\}.$$

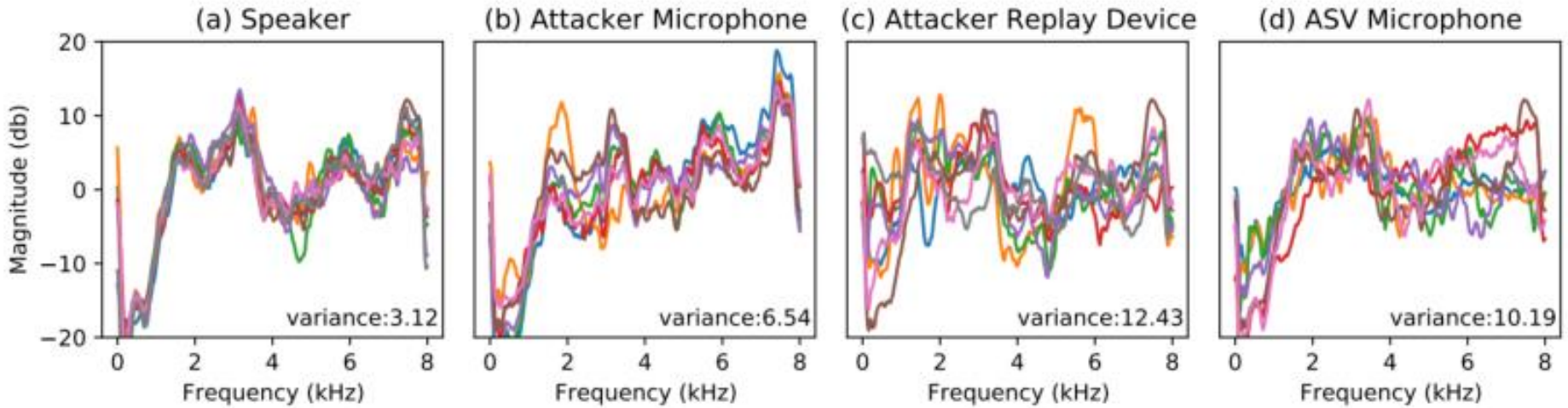
确定一个集合作为 锚点配置，在此基础上，每次只改变一个因素，观察H变化。



扬声器声场



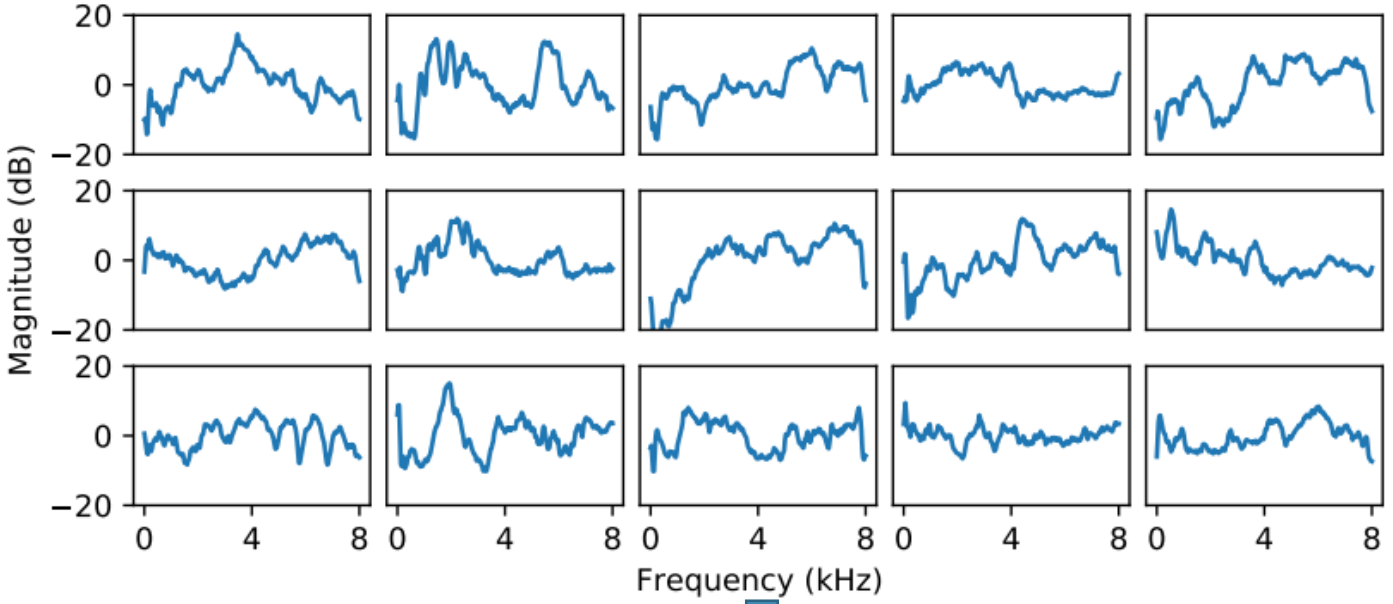
因子影响定量描述



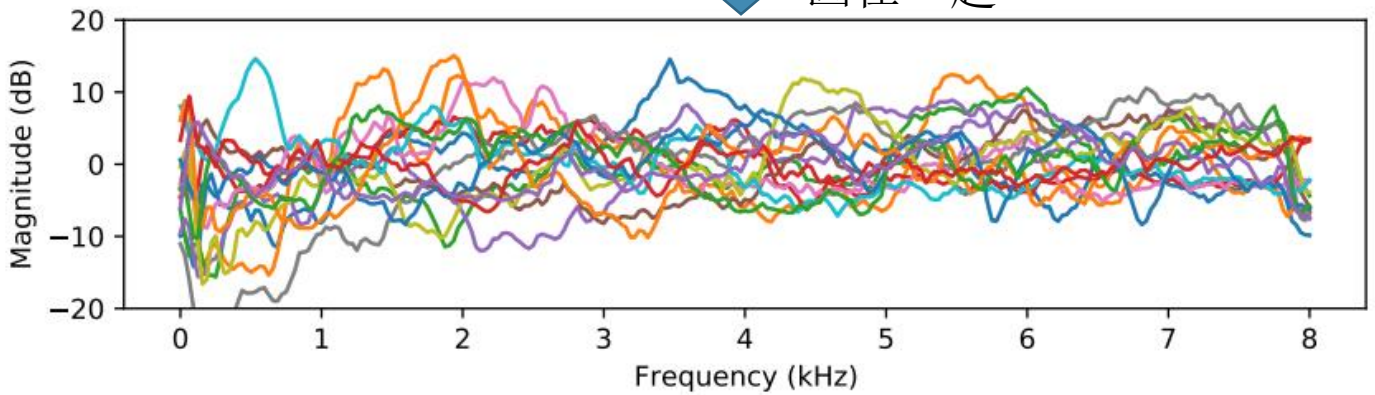
Factor	Full-bank Variance
Speaker	2.29 ± 0.92
Att Mic.	4.64 ± 1.81
Att. Replay	13.39 ± 4.67
ASV Mic.	9.75 ± 4.11

遍历所有可能的 锚点配置，计算每个锚点配置下variance的分布

传递函数的幅值观察



画在一起

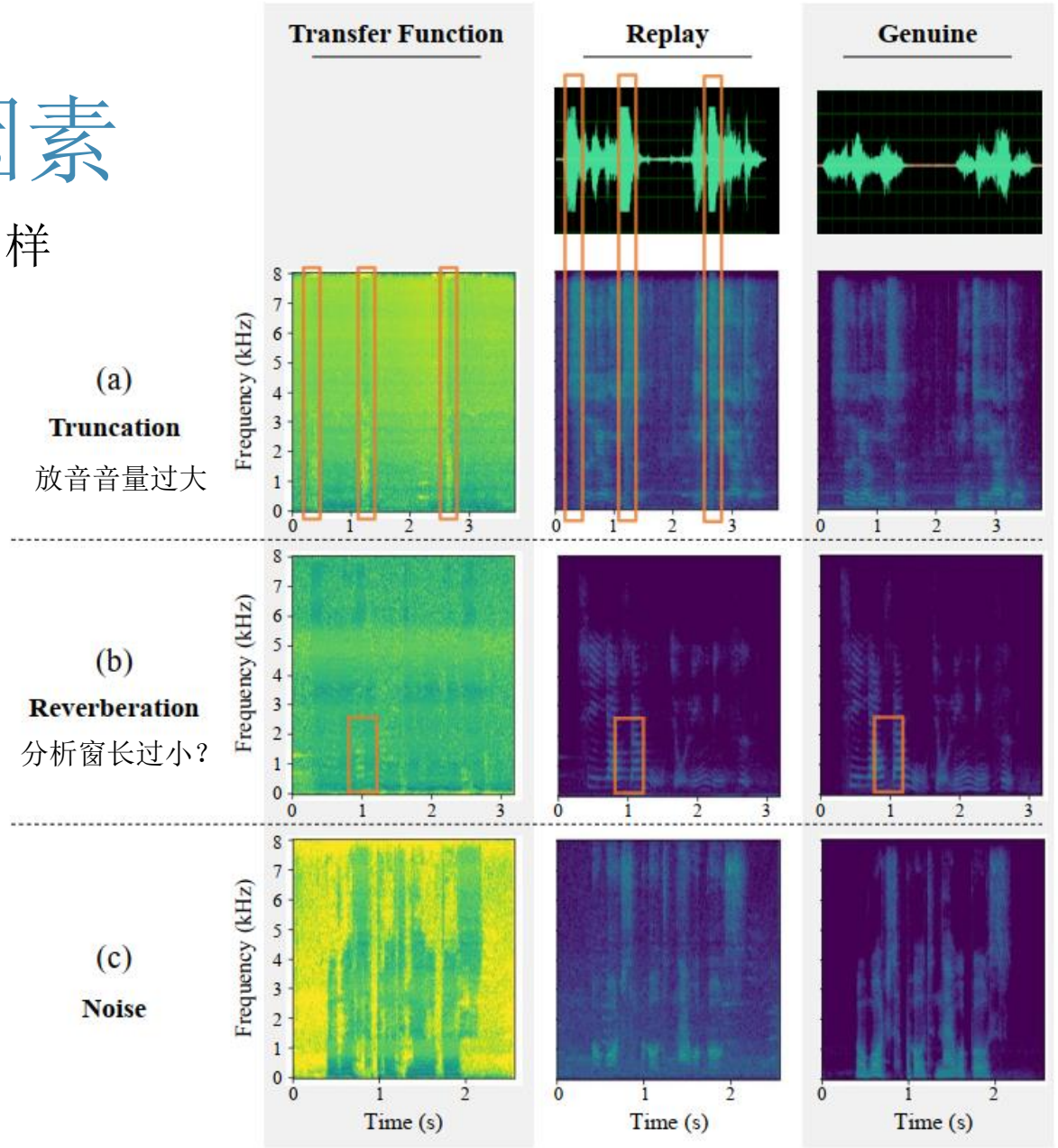


没有共性

难以（或不能？）预测

额外扰动因素

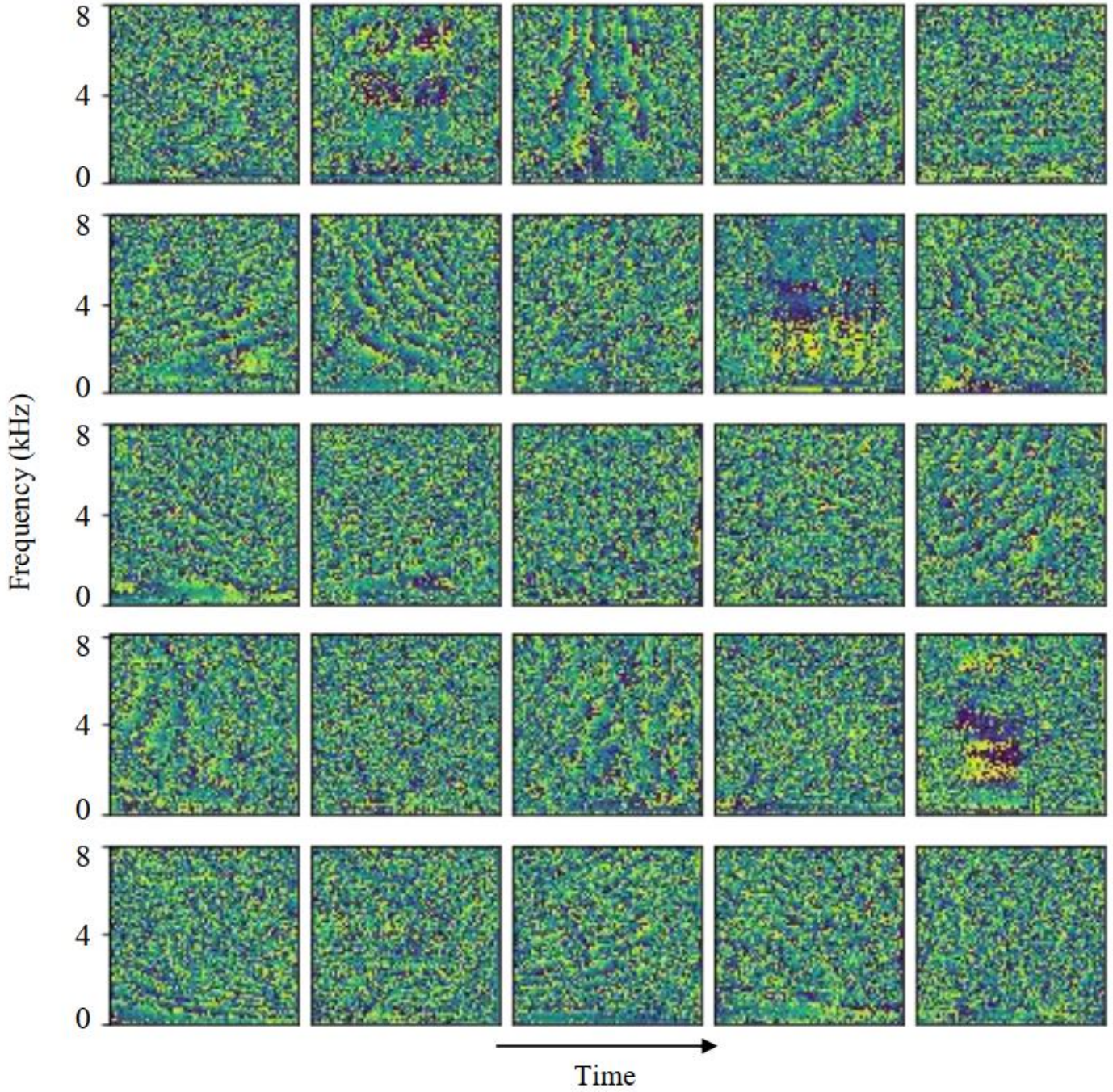
将导致H的变化更加多样



传递函数的相位观察

结构非常复杂
且存在**时变现象**

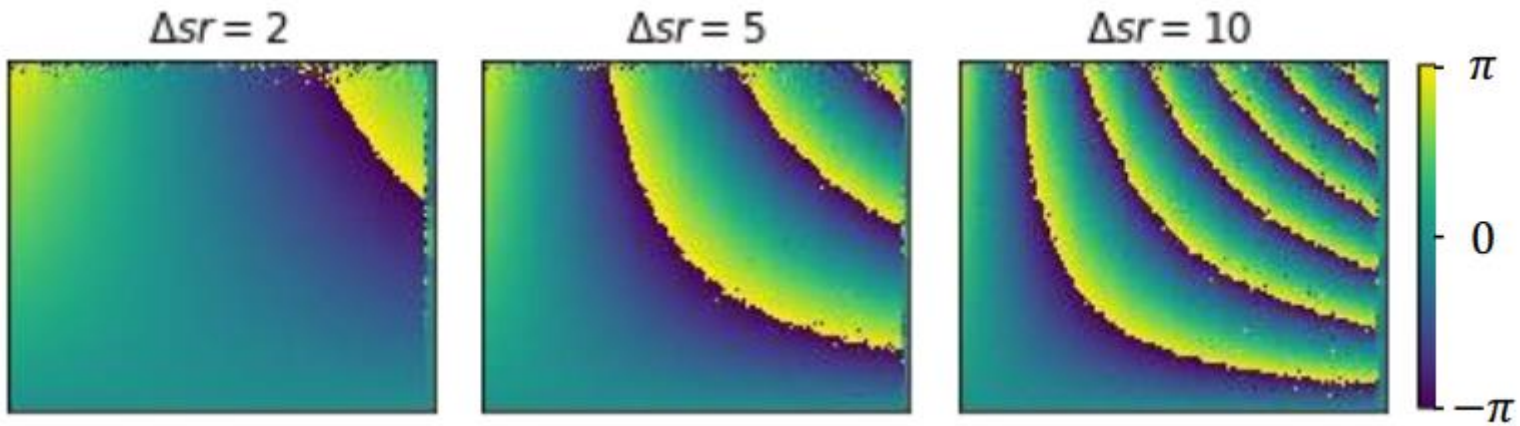
难以提取“代表性的”特征



时钟不同步造成的影响

由于麦克风、扬声器等设备的时钟不同步，将在相位上造成额外影响。

这使得相位更加复杂



针对重放的建模是困难的

数据分布分析

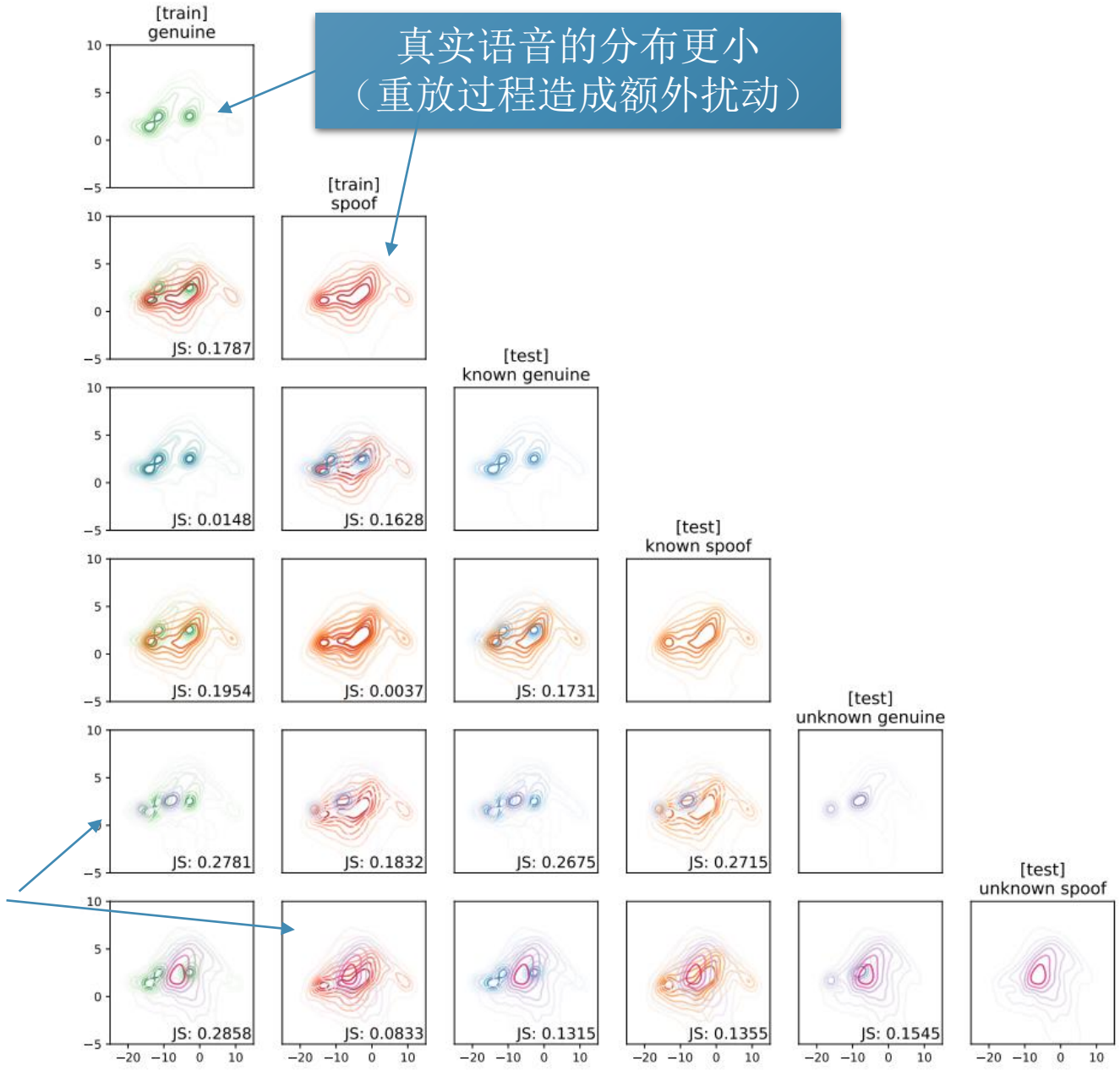
PRAD数据集划分

- 将部分因子划分在训练集，作为已知因子。剩余因子作为未知因子。
- 训练集：
  - 5 speakers,4 Att Mic, 5 Att Replay, 4 ASV Mic
- 测试集：
  - 包含全部因子，至多只有一个因子是未知的。

Table 4: Dataset distribution

Dataset	Condition	#genuine	#spoof
Train	/	100	800
Test	All known	100	1200
	Unknown speaker	60	720
	Unknown Att Mic	0	900
	Unknown Att Replay	0	720
	Unknown ASV Mic	75	900

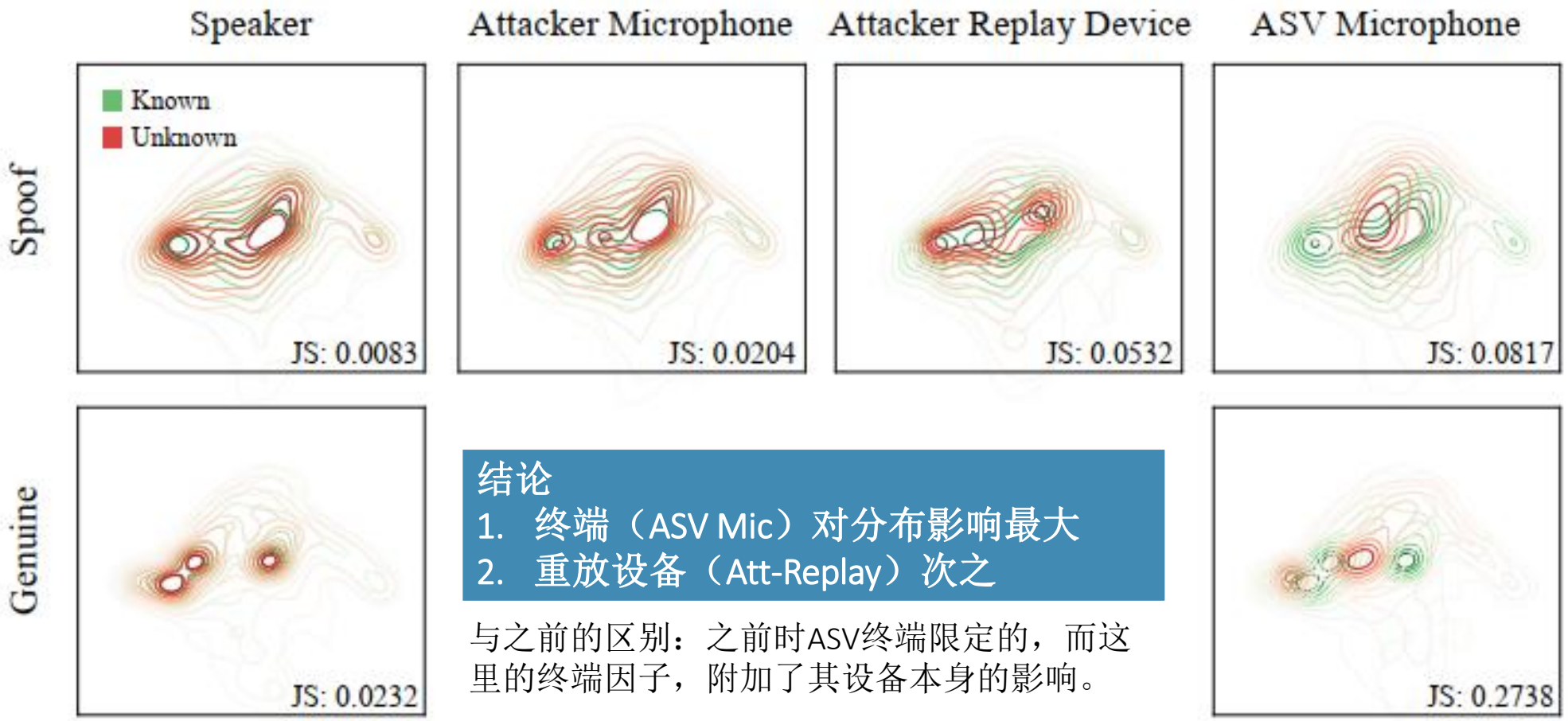
分布差异



针对终端因子的分析

不同因子所造成的分布差异

Type	Condition	KL Divergence	JS Divergence
Spoof	All known	0.98	0.17
	Unknown speaker	2.20	0.27
	Unknown Att Mic	2.70	0.36
	Unknown Att Replay	5.21	0.51
	Unknown ASV Mic	5.64	0.51
Genuine	All known	1.75	0.29
	Unknown speaker	3.61	0.40
	Unknown ASV Mic	9.25	0.60



结论

1. 终端（ASV Mic）对分布影响最大
2. 重放设备（Att-Replay）次之

与之前的区别：之前时ASV终端限定的，而这里的终端因子，附加了其设备本身的影响。

区分性分析结果

Table 6: Performance under different conditions. The EER threshold of 'all known' condition is used for FAR and FRR calculation. B1: CQCC-GMM; B2:LFCC-GMM;

Condition	FRR(%)		FAR(%)	
	B1	B2	B1	B2
All known	0.00	0.00	0.75	0.17
Unknown speaker	0.00	0.00	0.15	0.42
Unknown Att Mic	/	/	2.33	2.00
Unknown Att Replay	/	/	13.89	12.36
Unknown ASV Mic	69.33	84.00	2.67	3.44

两个麦克未知，对性能的影响是相似的

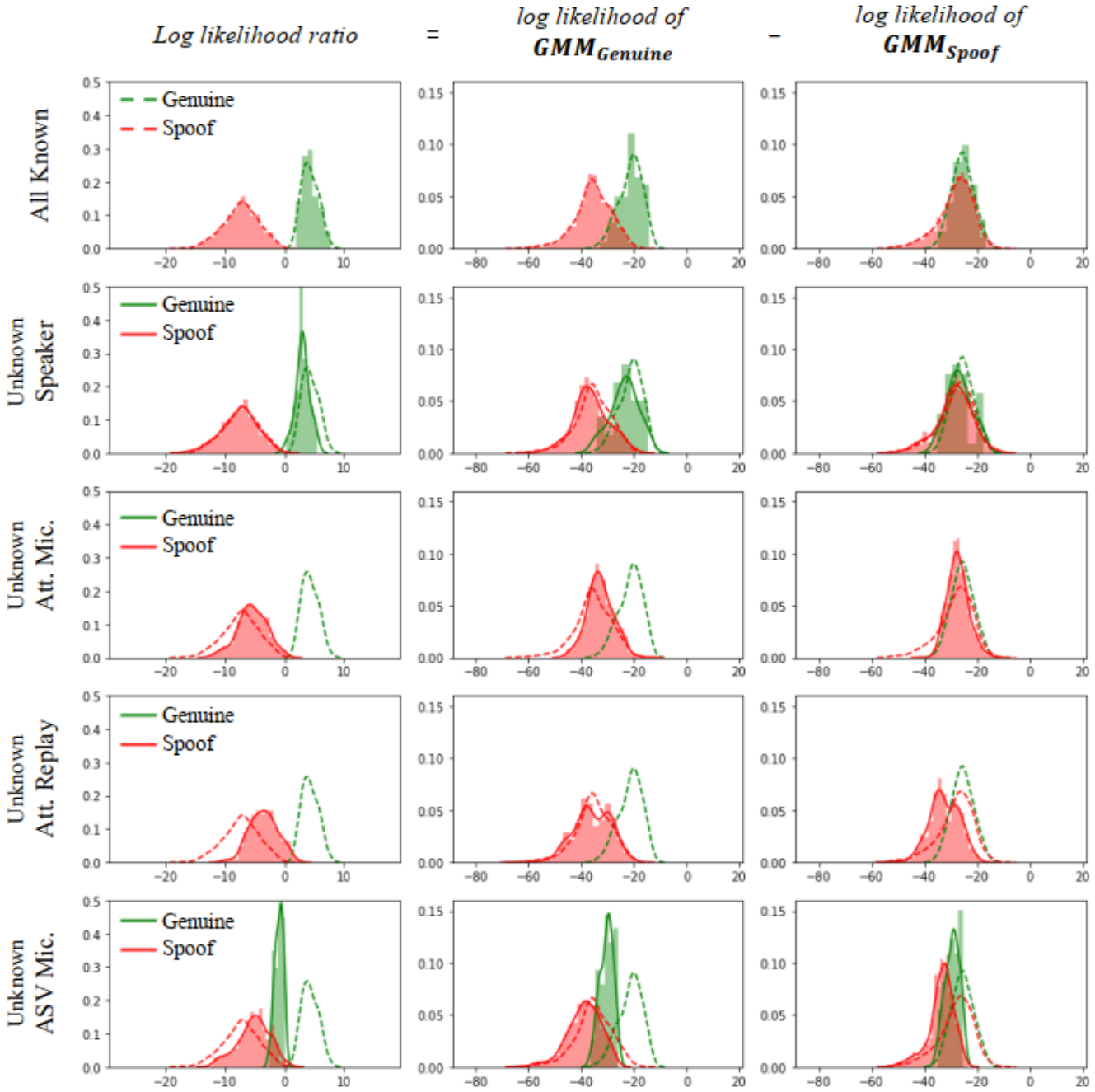
未知的重放设备将造成较多的闯入

终端改变时，真实语音的FRR极具上升。

分数分析

Speaker基本不影响分数

未知终端，导致**真实语音分数下降严重**

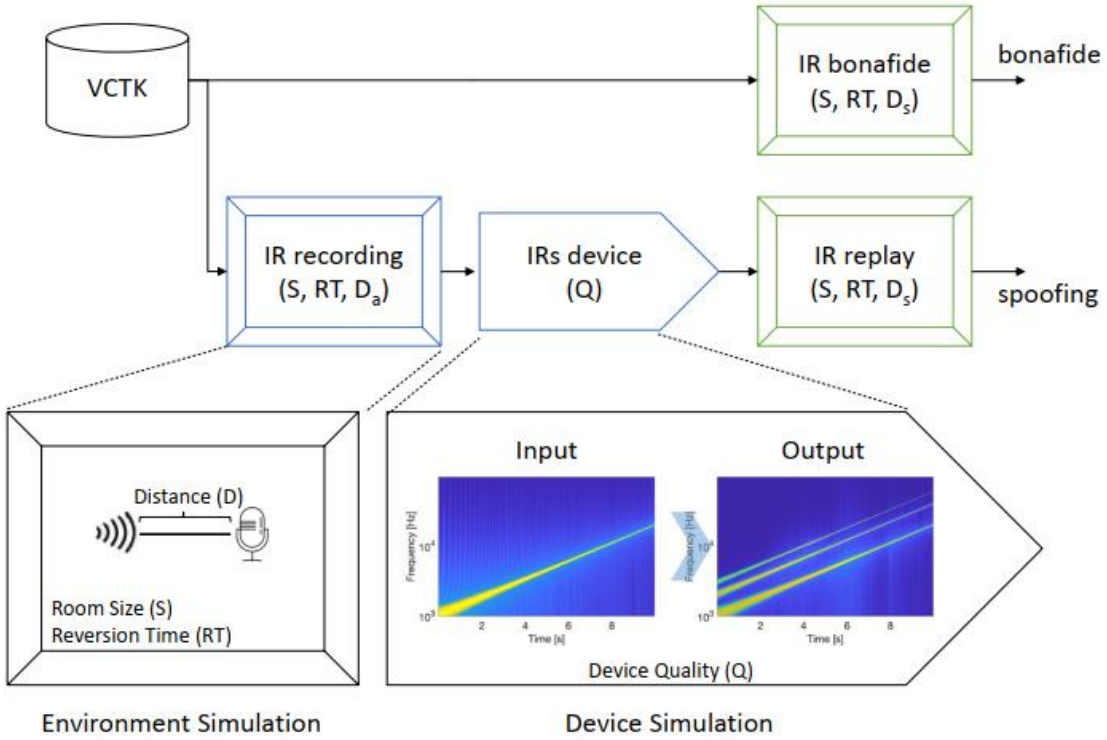


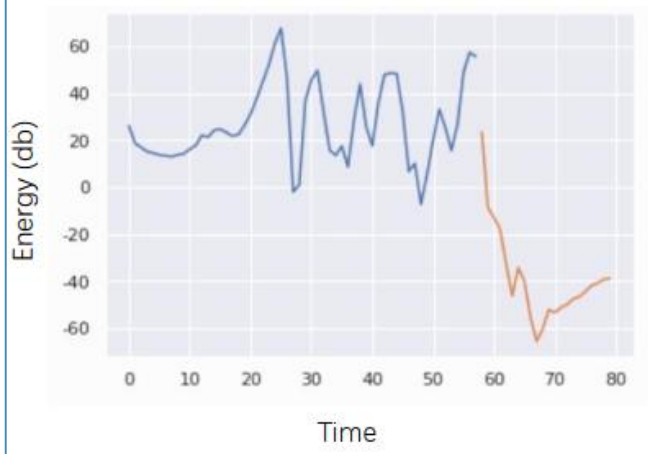
Spoof GMM似乎没有区分性

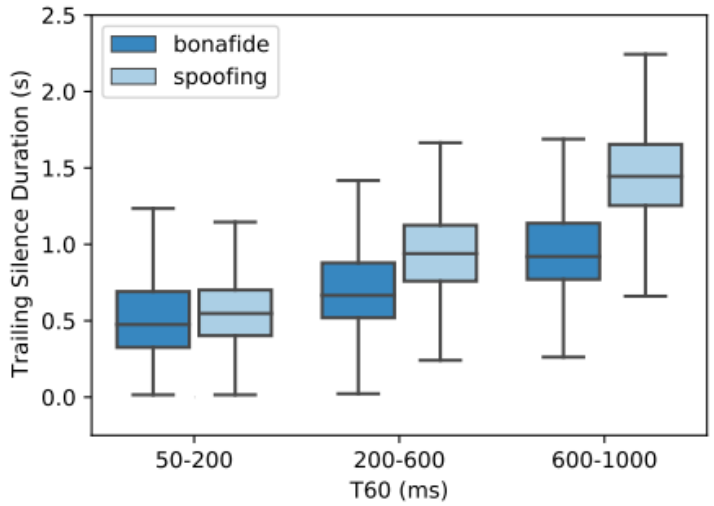
未知的重放设备将导致Spoof GMM分数明显降低

数据集的分析

ASVspoof 2019 数据库





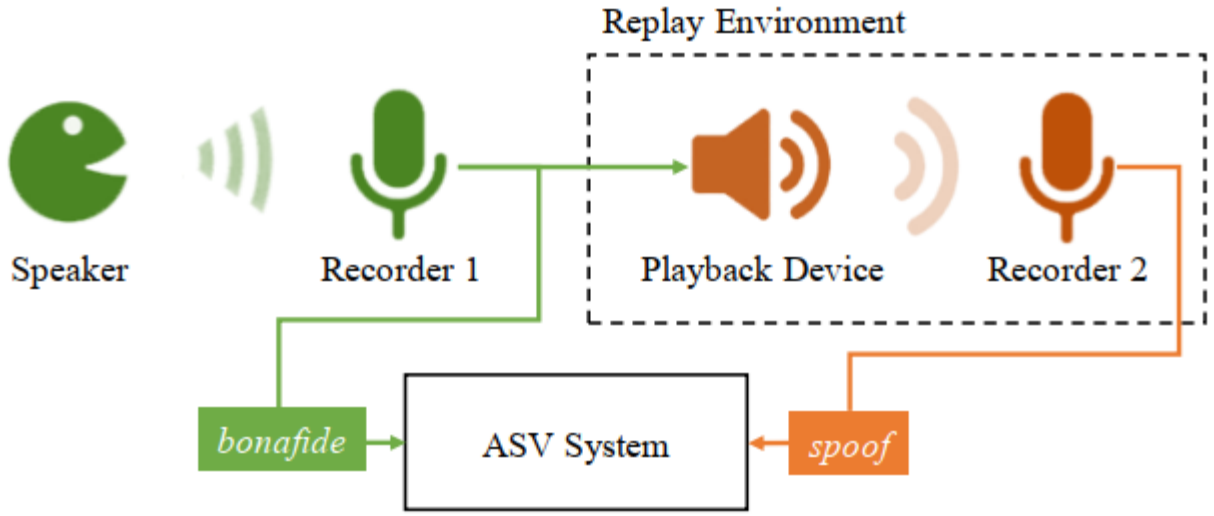


System	Condition			
	$O \to O$	$O \to R$	$R \to R$	$R \to O$
CQCC+GMM	9.87	15.39	14.33	9.90
Spectrogram+NN	3.15	15.05	3.59	4.33
CQTgram+NN	0.39	10.18	**1.13**	3.80
MGD+NN	0.97	12.66	2.45	**2.54**
CQTMGD+NN	**0.54**	**8.94**	1.36	3.61

O：原始；R：尾部静音被去除。

A->B：A上训练，B上测试

ASVspoof 2017 数据库



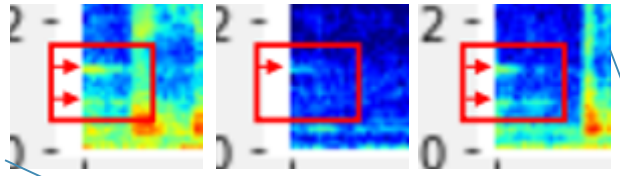
Subset	#Speaker	#Replay Configuration	#Bonafide	#Spoof
Train	10	3 (2 env, 3 pb, 1 rec)	1,507	1,507
Dev	8	10 (6 env, 6 pb, 7 rec)	760	950
Eval	24	57 (24 env, 23 pb, 24 rec)	1,298	12,008
Total	42	62 (26 env, 26 pb, 25 rec)	3,565	14,465

Env：环境；pb：重放设备；rec：录音设备，即Recorder2

ASVspoof 2017 数据库

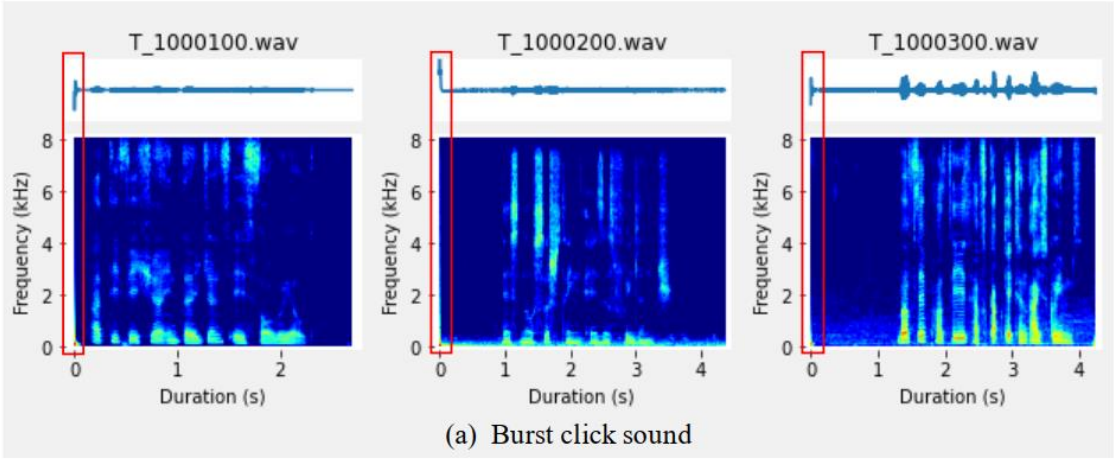
出现以下异常现象：

- 短促点击声音（Burst click sound, BCS）
- 全0静音
- DTMF（ Dual-tone multi-frequency signaling）声
- 真实语音前面有更长的静音

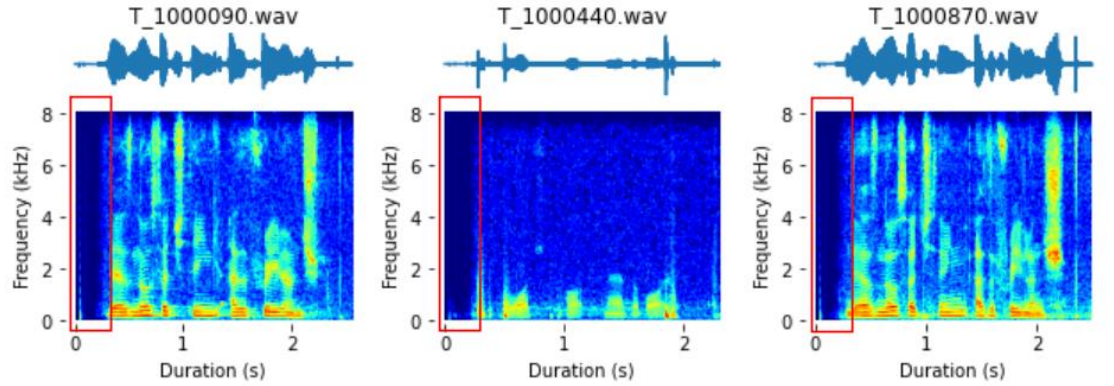


异常出现比例

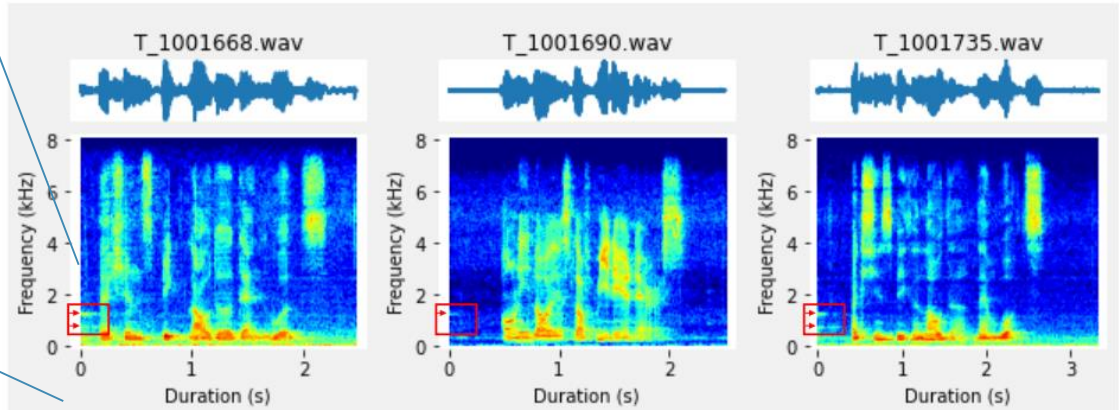
Artifact	Train		Dev	
	bonafide	spoof	bonafide	spoof
Burst click sound	36.36%	2.45%	41.06%	0.00%
Zero-valued silence	19.11%	0.00%	1.97%	0.00%
DTMF sound	0.00%	45.58%	0.00%	16.63%
Non-speech in first 300ms	60.45%	31.26%	73.55%	58.95%



(a) Burst click sound



(b) Zero valued silence



(c) DTMF artifacts

跨数据集结果

Table 10: Cross-dataset performance (EER%). Models are trained on the training subset and test on the evaluation subset. Since the AS19Real only contains an evaluation subset, we use the whole database either for training or for testing. (AS17: ASVspoof 2017[16]; AS19: ASVspoof 2019 PA[17]; AS19Real: ASVspoof 2019 Real PA[17])

System	Train	Test			
		PRAD	AS17	AS19	AS19Real
LFCC-GMM	PRAD	12.86	35.14	58.78	35.19
	AS17	45.71	33.91	81.22	47.59
	AS19	27.50	33.20	13.52	29.05
	AS19Real	30.72	35.52	20.10	/
CQCC-GMM	PRAD	14.31	37.98	36.76	26.83
	AS17	45.71	29.66	59.95	49.26
	AS19	29.28	37.98	11.25	12.57
	AS19Real	46.78	32.27	44.02	/

分布差异将导致跨数据集失效

结论

重放设备的传递函数不具有明显共性，难以建模和预测

基于生成模型的重放检测方法中，真实语音的模型比重放模型更加重要。

因此，这暗示可能one-class是一种较好的检测方案。

而不同数据集的分布不同，这将造成建模困难

此外，要小心数据集中存在的“偶然”特征

建议

- **对真实语音分布加以限制**的情况下，进行重放检测。
  - 例如：终端设备固定，限制真实语音的噪声较小等。
  - 理由
    - 重放设备特性难以（或无法）预测：不同设备的重放传递函数之间似乎没有共性
    - 重放检测更多依赖于真实语音的模型
    - 而当真实语音variance较大时，则将导致重放语音容易混入。
- 不应该包含训练集，避免**模型学习到数据集设计上的偏差**。