南京郵電大學
Nanjing University of Posts and Telecommunications

伪造语言检测及其泛化能力论文阅读

2025年4月24日

汇报人：朱家骏
指导老师：黄海平

01

Improving Generalization for AI-Synthesized Voice Detection

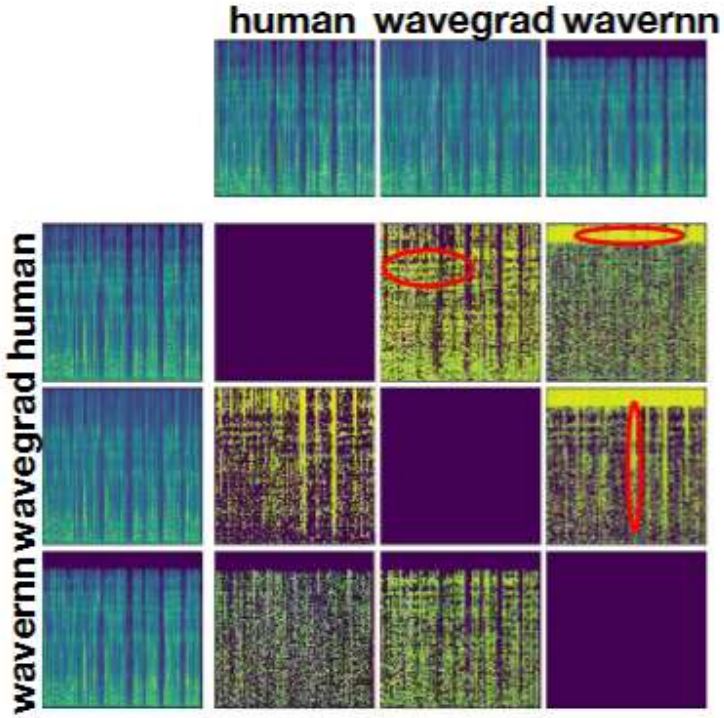
实验动机

到目前为止，针对音频深度伪造检测泛化问题的研究相对较少。以往的工作主要集中在域不变表示学习和自监督学习。然而，这些方法依赖于预定义的声码器，并且容易受到背景噪声、说话人身份等外部因素的影响。

（1）复杂的纠缠信息

人工智能合成语音检测中的泛化问题主要源于两个因素。

- 其一，许多检测器过度关注无关内容，如说话人身份和背景噪声
- 其二，不同的伪造技术会产生独特的伪影,如右图红色圈出地方



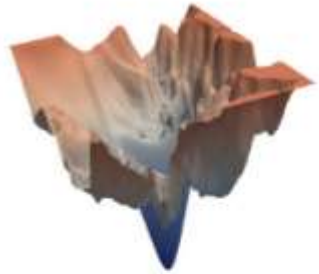
相同伪造声码器的数据在基线特征分布中紧密聚集，而不同声码器的数据则表现出更明显的分离。

实验动机

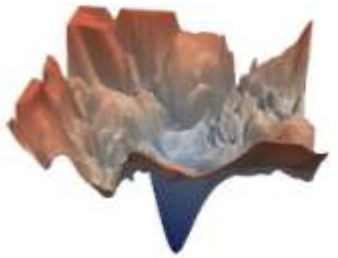
现有的基于深度神经网络的人工智能合成语音检测模型，如 RawNet2，参数化程度过高，在训练过程中容易记忆数据模式。这导致损失地形出现多个尖锐的极小值。这种尖锐性使得模型难以找到全局最优解，进而影响泛化能力。平坦化损失地形对于优化模型训练路径、增强泛化能力至关重要。

目标：

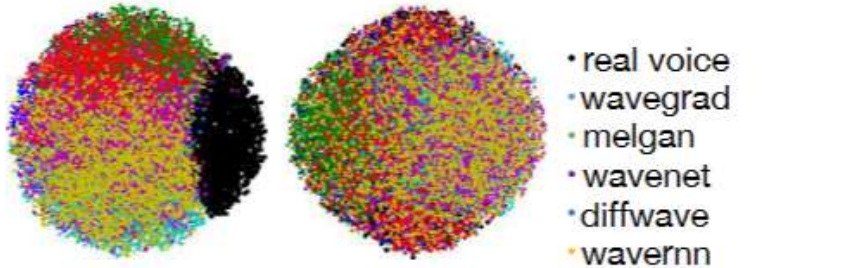
- 处理复杂纠缠信息：平等对待人工智能合成语音的域特征，并将其与人类语音特征区分开来。同时，对于人类语音和人工智能合成语音中的内容特征，检测器也应同等对待。
- 平坦化损失地形：优化模型训练路径、增强泛化能力。



(a) RawNet2



(b) Sun *et al.*



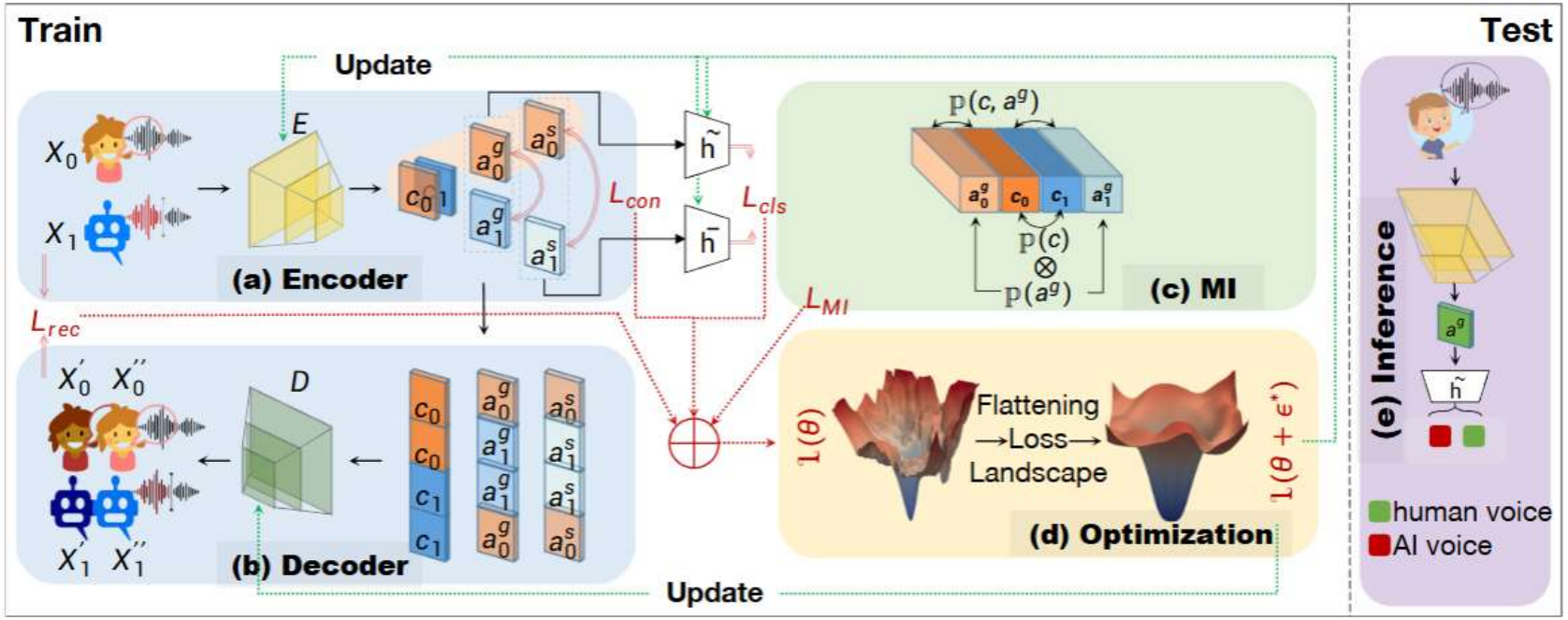
(d) Domain-agnostic (e) Content

02

方法概述

框架概述

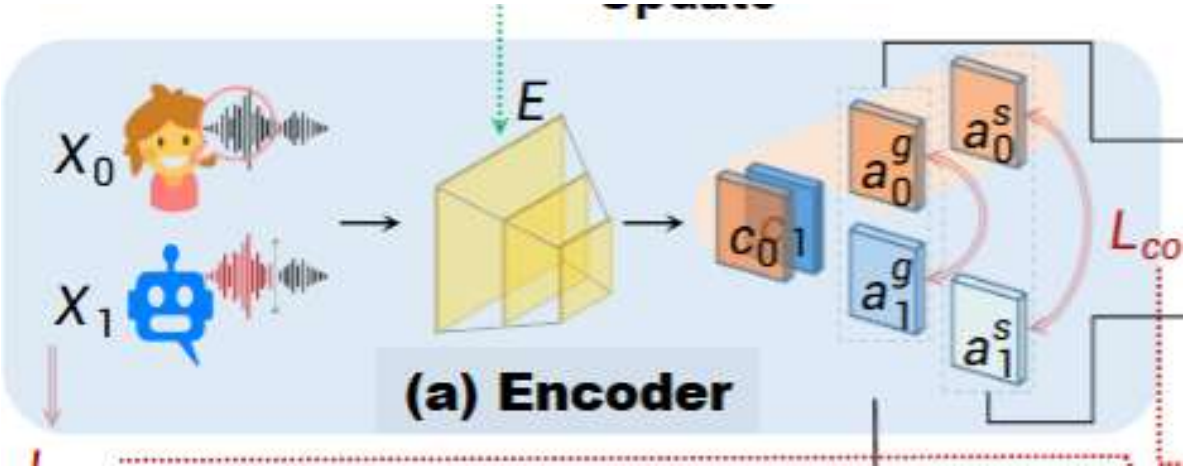
- 首先，引入一种新的解耦框架，结合多任务学习和对比学习，提取不同声码器共有的域无关伪影特征。
- 将域无关伪影与域特定伪影（即与特定声码器相关的伪影）分离，并应用重建正则化，确保原始语音和重建语音的一致性。
- 为使域无关特征具有通用性，进一步提升泛化能力，以内容特征分布为基准，应用互信息损失，使域无关特征与参考分布对齐。
- 最后，通过平坦化损失地形优化模型，避免陷入局部最优解，进一步增强泛化能力。



解耦过程

解耦学习模块：从输入语音中提取与声码器无关的伪影特征，用于检测。

一对语音(X_i,X_j)分别是合成人类语音或真实人类语音通过编码器的内容编码器和伪影编码器，提取内容特征c和伪影特征a，伪影特征包括域特定伪影a_s（特定声码器的特征）和域无关伪影a_g（不同声码器模型共有的特征）。



编码器的操作可表示为：$c_i, a_i^s, a_i^g = E(X_i)$。

分类损失与对比损失

分类损失：

利用多任务学习来分离域特定伪影和域无关伪影，对每个部分应用交叉熵损失。

损失函数可表示为：$L_{cls} = C(\bar{h}(a_i^s), D_i) + \lambda_1 C(\tilde{h}(a_i^g), Y_i)$ ，其中C(·,·)表示交叉熵损失，h^-和$h^\sim$分别是a_i^s和a_i^g的分类头，λ_1是超参数。

通过这种分类损失进行训练，编码器能够学习特定和共享的伪影信息，从而提高模型的泛化能力。

对比损失：

分类损失仅考虑了单个语音的信息，忽略了语音之间重要的全局相关性。利用铰链函数：

$$L_{con} = [b + \|a_{anchor} - a_+\|_2 - \|a_{anchor} - a_-\|_2]_+$$

其中a_{anchor}表示一个语音的锚点伪影特征，a_+和a_-分别是来自同一源的正样本和来自不同源的负样本的伪影特征。

- 对于域特定特征，对比损失促使编码器捕获特定声码器的表征；
- 对于域无关特征，鼓励编码器学习与特定声码器无关的可泛化表征。

重建损失

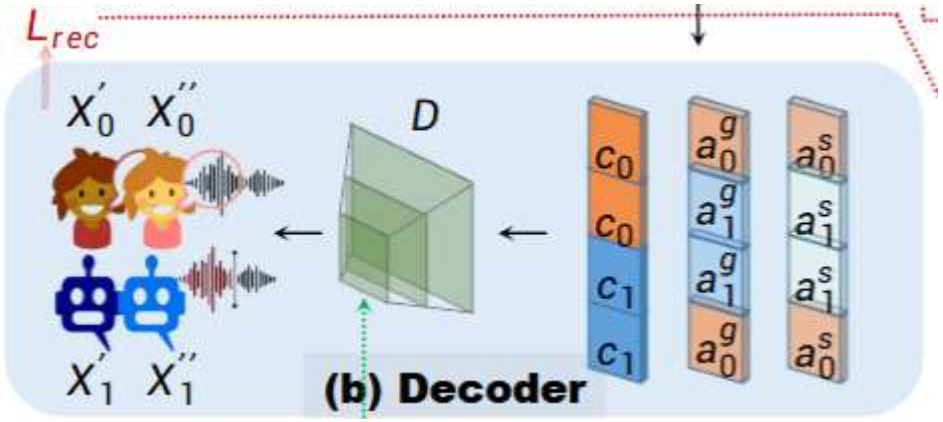
为确保提取特征的完整性，并保持原始语音和重建语音之间的一致性，应用重建损失：

$$L_{rec} = \|X_i - D(c_i, a_i^s, a_i^g)\|_1 + \|X_i - D(c_i, a_j^s, a_j^g)\|_1$$

其中D(·,·,·)表示基于解耦特征表示进行语音重建的解码器。

在L_{rec}损失中：

- 第一项是自重建损失，利用输入语音的潜在特征最小化重建误差；
- 第二项是交叉重建损失，使用配对语音的伪造特征惩罚重建误差。这两项共同促进特征解耦。

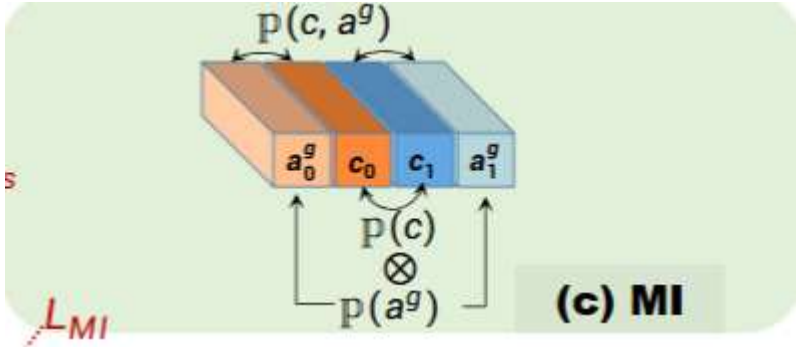


互信息损失与平坦损失地形的优化

互信息损失：

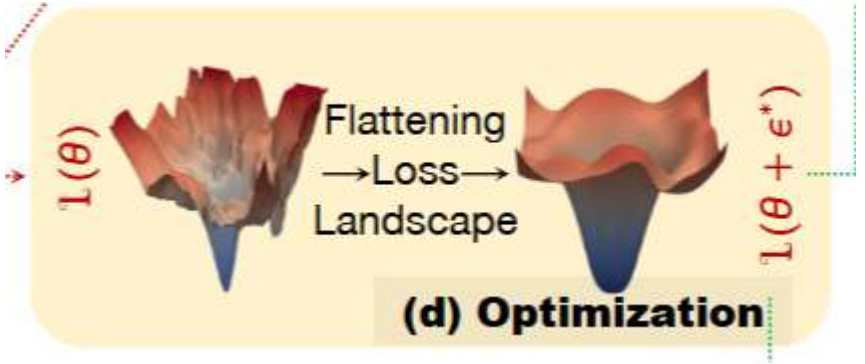
使用 KL散度来量化c和a_g之间的依赖关系。

通过使域无关特征与内容特征的分布相互对齐，可以直接实现保持不同声码器的域无关特征分布，从而增强泛化能力。



平坦损失地形的优化：

为了帮助模型避免在参数过多的深度神经网络中常见的次优解，并进一步提高泛化能力，我们应用了SAM（Sharpness-Aware Minimization）技术来平坦损失地形。



03

实验结果分析

实验结果

数据集：LibriSeVoc、WaveFake、ASVspoof 2019（以及FakeAVCeleb的音频片段

评估指标：错误率（EER）

编码器：RawNet2，但不包括最后一个全连接层

Methods	LibriSeVoc								ASVspoof2019							
	Seen vocoder				Unseen vocoder				Seen vocoder			Unseen vocoder				
	Avg	LSV	ASP	WF	Avg	ASP	FAVC	WF	Avg	ASP	LSV	Avg	ASP	FAVC	LSV	WF
LCNN (Lavrentyeva et al. 2019)	34.04	7.80	46.21	48.10	45.08	41.90	49.28	44.06	33.02	16.15	49.88	41.21	9.81	50.98	51.93	**52.13**
RawNet2 (Tak et al. 2021)	20.21	1.59	29.86	29.18	30.16	24.09	33.92	32.47	20.79	1.89	39.68	39.55	6.46	51.37	47.71	52.65
WavLM (Chen et al. 2022)	27.26	14.12	32.88	34.79	29.54	27.18	25.64	35.80	50.75	14.22	87.28	59.27	7.91	83.84	87.28	58.07
XLS-R (Babu et al. 2021)	33.47	11.21	45.37	43.83	45.11	51.23	42.91	41.18	53.65	9.04	98.26	74.85	6.40	94.91	98.26	99.82
Sun *et al.* (Sun et al. 2023)	18.67	3.79	22.77	29.45	27.86	24.42	25.52	33.65	22.18	3.92	40.44	41.59	8.38	54.78	50.59	52.62
Ours	**13.55**	**0.30**	**15.66**	**24.69**	**20.27**	**16.29**	**18.02**	**26.50**	**20.39**	**1.55**	**39.23**	**38.20**	**5.72**	**48.43**	**46.42**	52.24

在同域和跨域场景中均优于基线方法

01

Generalization Of Audio Deepfake Detection

实验动机

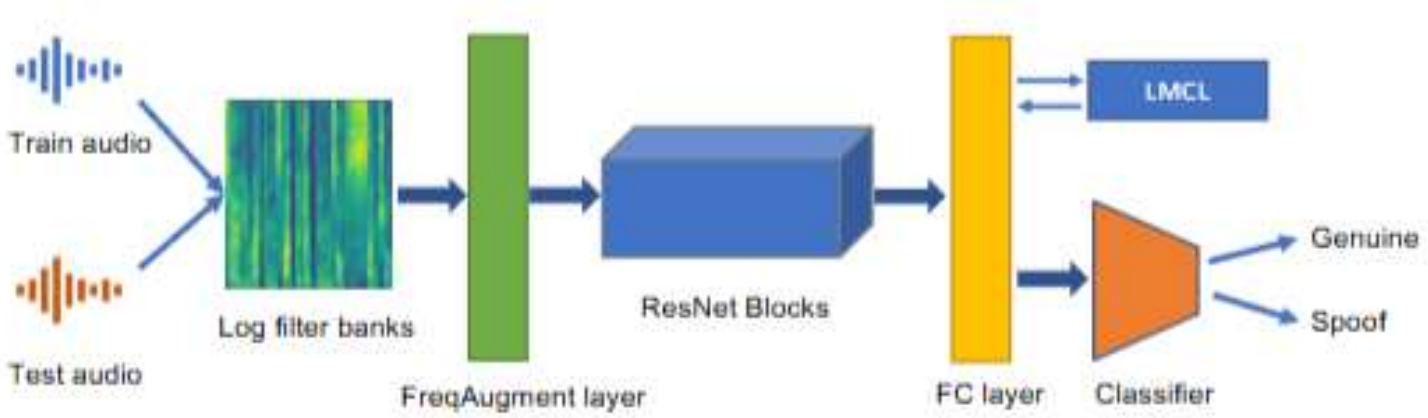
ASVspoof挑战赛的结果表明当前欺骗检测系统最大的问题在于其泛化能力。研究员现有的思路包括设计不同的低级频谱 - 时间特征，使用深度学习模型学习用于音频欺骗检测的判别特征嵌入，还有对不同传统声学特征和自动编码器学习到的特征进行了综合研究等，但效果都有局限性。因此作者从不同角度应对这一挑战，不再研究不同的低级音频特征，而是尝试提高模型自身的泛化能力。

02

方法概述

框架概述

- 使用大边际余弦损失函数（LMCL）。LMCL 的目标是最大化真实类和欺骗类之间的差异，同时最小化类内差异。
- 添加 FreqAugment 层。该层在 DNN 训练期间随机屏蔽相邻的频率通道，进一步提高 DNN 模型的泛化能力。
- 研究音频增强技术的有效性。利用公开可用的噪声，包括免费的电影、电视节目、音乐、其他噪声和房间脉冲响应，对音频文件进行增强处理，在噪声场景下训练和评估系统。
- 研究了所提欺骗检测系统在呼叫中心环境中的性能。因此，通过 VoIP 信道对 ASVspoof 2019 数据集进行逻辑重放，模拟欺骗攻击。



从原始音频中提取 60 维线性滤波器组（LFBs），并输入到残差网络中。在训练 ResNet 模型时，使用 FreqAugment 层和大边际余弦损失。训练完该模型后，将相同的训练话语输入到 ResNet 中，提取欺骗嵌入，然后用于训练后端真实与欺骗分类器。

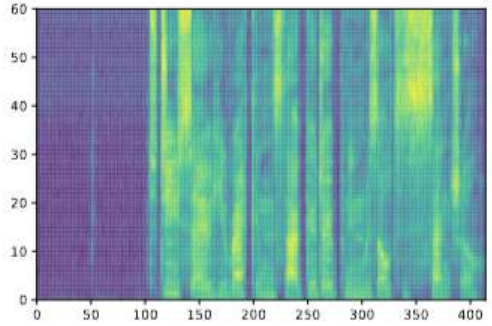
低级特征与频率掩蔽

低级特征：

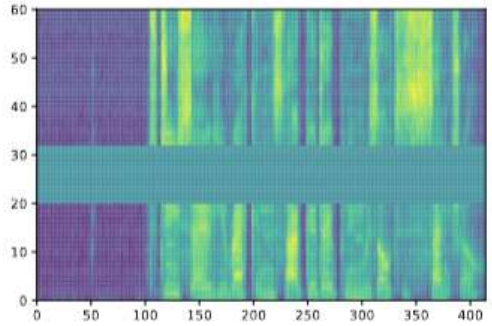
该系统中直接使用了线性滤波器组（LFBs）提取出的频谱能量特征，作为语音信号的基础输入特征，以捕捉原始音频的频率结构和伪造伪影。使用在 30ms 窗口上提取的 60 维 LFBs，帧移为 10ms。在话语级别进行均值和方差归一化。

频率掩蔽：

在线频率掩蔽是一种在训练过程中随机“遮挡”语音频谱中的某段频率区域的策略，它可以提升模型鲁棒性，防止过拟合于局部频段特征，广泛用于语音处理任务的增强手段。



(a) Linear filter banks of an audio signal of 4.1 seconds.



(b) Output of the FreqAugment layer.

原始 LFB 和掩蔽后 LFB 的对比。

大边际余弦损失与深度残差网络

大边际余弦损失：

是一种在分类任务中增强类间可分性的方法，它通过引入一个“角度间隔（margin）”来让模型学到更判别性更强的特征，拉大类间边界。

$$L_{lmc} = \frac{1}{N}\sum_i -log\frac{e^{s\left(cos\left(\theta_{y_i,i}\right)-m\right)}}{e^{s\left(cos\left(\theta_{y_i,i}\right)-m\right)}+\sum_{i\neq y_i} e^{scos\left(\theta_{j,i}\right)}}$$

残差网络（ResNet）：

一种深度神经网络结构，通过引入“跳跃连接（skip connection）”使网络直接学习输入与输出之间的残差，从而缓解深层网络中的梯度消失和性能退化问题。它允许信号在网络中跨层传播，使得非常深的模型（如ResNet-50、ResNet-101）也能高效训练并获得良好的泛化性能.

数据增强策略

使用了两种失真方式来增强 ASVspoof 2019 数据集：混响和背景噪声

用于**混响**的房间脉冲响应（RIR）选自公开可用的 RIR 数据集

背景噪声：

音乐、电视、嘈杂人声和免费音效

数据来源：MUSAN 噪声语料库、Youtube 上公开的电影和电视节目

具体实现：对音乐、电视与语音话语使用随机选择的 RIR 进行混响，再将混响后的噪声添加到混响后的语音话语中。对于嘈杂人声和免费音效噪声，先将背景噪声文件添加到干净音频中，然后使用随机选择的 RIR 对混合音频进行混响。添加噪声时的信噪比在 5dB 到 20dB 之间随机选择。

03

实验结果分析

实验结果

南京邮电大学 Nanjing University of Posts and Telecommunications

1.训练数据集相关

1. **T1**：是 ASVspoof 2019 LA 挑战的官方训练协议所使用的数据集，包含 25380 个语音样本，仅使用 ASVspoof 2019 的原始训练集。
2. **T2**：是在 T1 的基础上，添加了经过数据增强的训练集。数据增强采用修改自 Kaldi 的数据增强技术，通过添加混响和背景噪声（如音乐、电视、嘈杂人声、免费音效等）来扩充数据，总共包含 152280 个语音样本。
3. **T3**：在 T2 的基础上，加入了逻辑重放的训练集。利用 Twilio 的语音服务对 ASVSpoof 2019 数据进行语音通话重放并录制，模拟呼叫中心环境，该训练集采样率从 16kHz 降至 8kHz，再上采样回 16kHz 使用，样本总数达到 177660 个。

2.评估基准相关

1. **E1**：是 ASVspoof 2019 LA 挑战的官方评估集，包含 71237 个语音样本，用于评估模型在原始测试条件下的性能。
2. **E2**：是增强评估集，在 E1 的基础上进行了与 T2 中训练集相同方式的数据增强，即添加混响和背景噪声，样本数量扩展到 356185 个，用于评估模型在噪声环境下的性能表现。
3. **E3**：是逻辑重放评估集，同样利用 Twilio 的语音服务模拟呼叫中心环境，对数据进行逻辑重放后得到，用于评估模型在模拟呼叫中心场景下的性能，样本数为 71237 个。

	Benchmarks		
Training Protocols	E1	E2	E3
T1	1.81%	20.43%	8.70%
T2	1.64%	5.34%	8.21%
T3	**1.26%**	**5.32%**	**2.62%**

Model	Training protocol	EER	t-DCF
ResNet18	T1	4.04%	0.109
ResNet18-L	T1	3.49%	0.092
ResNet18-L-FM	T1	**1.81%**	0.052

用 LMCL 取代 softmax 后，等错误率降至 3.49%，t-DCF 降至 0.092。这表明 LMCL 能够迫使模型学习更具鲁棒性、泛化能力更好的特征。然后，添加频率掩蔽层，进一步将等错误率降至 1.81% 。利用音频增强技术后，等错误率从 1.81% 进一步降至 1.64%。逻辑重放后，等错误率从 1.64% 进一步降至 1.26%。

感谢观看