arXiv:2611.29871v1 [eess.AS] 23 Nov 2026

该信任哪一层？信号损坏条件下用于歌声深度伪造检测的冻结语音编码器层选择

*Benjamin Hurt**、*Yefei Yang**

加利福尼亚大学圣塔芭芭拉分校

摘要

冻结的自监督语音编码器已被广泛用于音频深度伪造检测。然而，层通常是从开发集中选择的。我们在 CtrSVDD 上研究了这一选择对歌声深度伪造检测的影响。我们在冻结的 XEUS（3–4 个随机种子）和 XLS-R（2 个随机种子）上，为每一层分别训练一个检测头，并额外训练一个学习式层混合。在干净音频以及噪声、混响和有损编码条件下进行测试。在干净音频上，对于两种编码器，中间层的表现都优于最后一层，而学习式混合并未超过最佳单层。开发集选出的层，其等错误率比最佳测试层分别差 2.3 个百分点（XEUS）和 4.1 个百分点（XLS-R）。在噪声和编码条件下，XEUS 的最佳层仍位于中间位置；而在混响条件下，其较高层的表现迎头赶上。相比之下，XLS-R 的最佳层向更浅层移动。在探索性分析中，XEUS 使用更早的层检测商业合成攻击时表现最佳，尤其是在噪声条件下。因此，仅凭域内开发集和干净音频无法可靠地选择层。

索引词— 歌声深度伪造检测、自监督学习、层选择、鲁棒性

1. 引言

自监督学习（SSL）编码器已被用作音频深度伪造检测的标准前端 [1, 2]。冻结编码器后，会在其隐藏状态上训练一个轻量级后端。例如，据报道，采用多头因子化注意力池化（MHFA）后端的冻结 XEUS，只需使用微调系统中一小部分可训练参数，便能在多个语音深度伪造基准上实现良好的泛化能力 [3]。然而，这类系统仍须决定使用哪个隐藏层，或使用哪些层的混合表示。逐层分析表明，语音自监督学习模型的不同深度编码着不同信息 [4], ；近期一项针对音频深度伪造检测的自监督学习模型研究发现，在不同语言和伪造类型（包括歌曲）中，较低层的判别性最强 [5]。基准测试通常通过对各层进行学习得到的加权和来回避这一选择问题 [6]。

这两个方面很少被放在一起考察。首先，研究者通常在干净条件下的测试音频上比较各层；而在部署中，检测器接收的音频往往会受到背景噪声、房间环境和有损编解码器的影响而劣化——在语音反欺骗领域，信道和编解码器效应已被确认为造成性能下降的一个来源 [7]。其次，与任何超参数一样，层通常是在包含训练攻击的开发集上选定的。在这些攻击上表现良好的层，可能无法泛化到未见过的攻击。

我们针对歌声深度伪造检测（SVDD）研究了这两个问题，因为面向语音构建的检测器无法直接迁移 [8,9] ，而最具挑战性的系统依赖自监督学习前端 [10]。

*贡献相同（共同第一作者）。

使用CtrSVDD基准 [9], ，我们针对冻结编码器的每一层以及学习式层混合分别训练检测头；对于XEUS [11], 每个随机种子采用同一套方案和同一数据顺序（每个条件使用3–4个随机种子、干净音频和九种音频损坏）；对于XLS-R [12] ，则使用2个随机种子、干净音频和三种音频损坏。我们发现：(i) 对于两种编码器，在干净音频上，中间层均优于最后一层，而学习式层混合并未优于最佳层；(ii) 开发集无法可靠地选出层：它几乎无法区分各个中间层，而其选择使主要等错误率分别损失2.3和4.1个百分点；(iii) 最佳深度会随信号条件变化，并且在不同编码器中的变化方式不同，在混响条件下尤为明显；(iv) 在探索性分析中，XEUS针对域外商业攻击的最佳层不同于针对其他攻击的最佳层，且在噪声条件下差异逐渐增大。

2. 实验设置

数据与指标。 CtrSVDD [9] 包含16 kHz单声道歌声片段，分为训练集（84,404）、开发集（43,625）和测试集（92,769）。训练集和开发集包含相同的攻击样本A01–A08；测试集使用A09–A14，其中没有任何样本出现在训练集中。A14由商业歌声合成器生成，并因其属于域外数据而未纳入挑战赛排名 [10]。我们以两种方式报告等错误率（EER；越低越好）：主要等错误率，在全部13,596个真实片段与53,983个A09–A13 欺骗样本之间计算；以及*A14* 等错误率，在相同的真实片段与25,190个A14 欺骗样本之间计算。与官方挑战赛排名一样，主要等错误率不包括A14；但与官方排名不同，我们的两个指标都保留ACESinger真实片段，而排名将其排除 [10]。每个音频片段均截取其前64,600个采样点，或用零填充至该长度（约4秒）。

编码器和头。 XEUS [11]是一个拥有 19 层的 E-Branchformer编码器，通过掩码预测、去噪和去混响，在 4,057 种语言的逾百万小时语音数据上进行预训练，对
目标。XLS-R 300M [12] 是一个 24 层的 wav2vec 2.0 Transformer，在 128 种语言的 436k 小时数据上进行预训练，未采用去噪或去混响目标。我们将两个编码器都保持冻结。仅训练头、融合模块和混合权重。我们提取每一层的输出；对于两个编码器，最后一层均取自最终层归一化之后的输出。[1] 每个单层头依次应用线性投影和层归一化、多头因子化注意力池化（MHFA） [13] （K=64, H=8）、第二次层归一化以及线性分类器。学习混合头对所有层的输出进行池化，使用经过 softmax 加权的和，初始权重设为均匀分布，并采用相同的训练方案从头开始训练。仅对于 XEUS，我们还会在 [14], 的自监督学习–CQCC交叉注意力融合模块中训练相同的头，该模块加入了恒Q倒谱特征。

[1]ESPnet 的 XEUS 前端会对整个批次的输入波形进行标准化，因此某个音频片段的得分会受同一批次中其他音频片段的影响。我们对每个音频片段分别进行标准化。在此流水线较早的检查点上，这使等错误率变化了 0.15–0.18 个百分点（主结果）和 1.1–1.4 个百分点（A14）。

训练。 一个随机种子的所有头均在同一项作业中训练。每个批次只经过一次编码器，然后传递给所有头。因此，同一随机种子的各个头会以相同顺序看到相同数据，同时每个头保留自己的优化器、梯度裁剪和早停设置，从而使同一随机种子内不同层之间的差异具有配对性。各头使用 AdamW（学习率 10^{-4}，权重衰减 10^{-4}；10^{-3} 混合权重不采用权重衰减）、带两个预热周期的余弦调度、最多 20 个周期、批量大小 24 以及类别加权交叉熵。每个头保留开发集等错误率最低的检查点，并在连续六个周期未取得改进后停止训练。我们为 XEUS 训练四个随机种子，为 XLS-R 训练两个随机种子。

测试时损坏。仅对测试音频应用损坏；每个音频片段采用确定性处理，并在截断前进行；训练仅使用干净音频。加性噪声来自 MUSAN 噪声子集 [15]，信噪比分别为 20 、 10 、 5 和 0 dB，按完整音频片段测量，因此评分所用的 4 秒片段的信噪比会有所变化。混响处理将每个音频片段与 RIRS NOISES 集合 [16]（RWCP、AIR 和 REVERB 2014）中的 325 个真实房间脉冲响应之一进行卷积，并恢复输入电平。有损编码采用 128 和 32 kbit/s 的 MP3，以及 32 和 12 kbit/s 的 Opus，解码后还原为 16 kHz。对于 XEUS，四个随机种子均在干净音频、5 和 0 dB 噪声以及混响条件下进行评分，其他六种条件则使用三个随机种子。XLS-R 在干净音频、5 和 0 dB 噪声以及混响条件下进行评分。

分析。 我们报告随机种子均值和标准差，并通过逐随机种子配对差值以及符号一致的随机种子数量来比较各层。差值以百分点（pp）表示。为比较不同深度的编码器，我们使用相对深度（层索引除以层数）。*开发集选定层* 的开发集等错误率均值最低，这是我们预先确定的规则。*最佳测试层* 的测试等错误率均值最低；该层是在事后选定的，仅作为参考。由于 A09–A13 欺骗样本仅来自 51 名测试集歌手中的 11 名，我们还通过歌手重采样检查关键差异是否仍然存在：在测试集歌手上进行按类别分层的 bootstrap（B=1000，在各头之间配对）可得到 95% 区间；表1中的星号表示其区间不包含零的差异。A14 分析（第 3.4 节）并非预先计划：我们在查看三个 XEUS 随机种子后选择了相关比较，而之后训练的第四个随机种子则作为留出检验。我们还在一半测试片段上选择最佳 A14 层，并测试

另一方面也如此，两个方向均进行测试（按剪辑 ID 将数据分成两半，因此歌手有所重叠）。

3. 结果

3.1. 干净音频条件下的 XEUS

图 1 展示了干净条件下的结果。主要等错误率从 L1 的 16.9% 下降，在 L9–L12 达到最低值，随后向顶层回升。在全部四个随机种子上，最后一层的表现都不如 L10，主要等错误率高出 1.99 个百分点，A14 高出 4.80 个百分点。只有 A14 上的差距在歌手重采样后仍然存在。四个随机种子不足以区分 L 9 12 10 – 11 −0.57 ± –L 之间的层：0.77 10, 12, 11, 9 L L 为 pp，而且每个随机种子的最佳层都不同（L L L L）。学习得到的混合权重主要集中在 L6–L12，在 L19 上几乎没有权重（0.014）。它不如 L10（8.84% 对比 7.50%；4 个随机种子中有 3 个），在主要等错误率上则优于开发集选定的 L8（低 0.99 个百分点），但在 A14 上不如 L8（高 1.57 个百分点），在噪声条件下也不如 L8。加入 CQCC 融合后，结果特征相同，且没有稳定的增益（融合 L10：7.42%）。

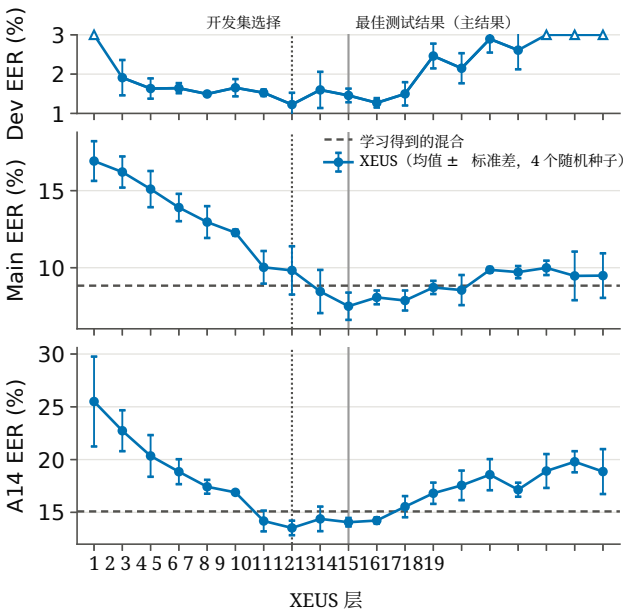

图 1。仅使用 XEUS 头，在干净测试集上的各层开发集、主要及 A14 等错误率（四个随机种子的均值 ± 标准差）。虚线：学习式层混合。竖线：开发集选定层（L8）和主要等错误率下的最佳测试层（L10）。三角形：开发集等错误率超出坐标轴范围。

3.2. 在开发集上选择层

开发集等错误率几乎无法区分中间层。对于 XEUS，L3–L12 的开发集等错误率仅在 1.23–1.65% 之间，而相同层范围内的主要测试等错误率则在 7.50–15.10% 之间。对于全部 19 层，开发集和测试集的排名相关性较弱（Spearman $\rho = 0.29$, $p = 0.24$ 0 04 0 28 8 13 ；每个随机种子为 . – .），而在 L –L 范围内则完全不相关（ $\rho =$ −0.03 8, 10）。使用四个随机种子时，该规则选出 L，而它在全部四个随机种子上的表现都不如 L，平均高出 2.32 个百分点；歌手重采样后，这一差距仍然存在。使用三个随机种子时，该规则选出 L11。对于这四个随机种子的全部 15 种子集组合，该规则选出 L8 十次、L11 三次、L9 两次，通常开发集等错误率的差距极小（中位数为 0.056 个百分点），而且从未选出最佳测试层（平均损失 1.76 个百分点）。学习得到的混合在所有头中开发集等错误率最低（0.92%），但其主要测试等错误率排名为第 20 名中的第七名。XLS-R 的问题更为明显：该规则在两个随机种子上都选出 L4，其干净条件下的主要等错误率比最佳测试层（L9）高 4.13 个百分点，且 $\rho = 0.36$ ($p = 0.08$)。在 A14 上，两个编码器的情况不同。对于 XEUS，开发集与 A14 的排名一致（$\rho = 0.74$），开发集选定的 L8 也是干净条件下 A14 表现最佳的层（第 3.4 节）；对于 XLS-R，两者则不一致（$\rho = 0.12$）。

3.3. 损坏条件下的 XEUS

表1总结了 XEUS 在损坏条件下的结果，图2展示了每一层的等错误率与各条件下最佳层等错误率之差。在干净音频以及所有噪声和编解码器条件下，最佳层均位于 L9–L12，30 个随机种子对应的最佳层中也有 28 个位于此范围内。噪声会损害每一层的性能（在 20 dB 下，L10 从 7.50% 上升至 15.66%），但对上层的影响最大：L19 – L10 相较于干净音频的差距从 1.99 百分点扩大至 4.2–7.3 百分点，且在每个随机种子上 L10 均表现更好。128 kbit/s 的 MP3 几乎不改变这一分布，而 12 kbit/s 的 Opus 会损害

表1。XEUS 单独使用的头在损坏条件下的测试等错误率（%）。 n：随机种子数。最佳层：事后选取种子平均测试等错误率最低的层。差值：各随机种子的配对差值的平均值，单位为百分点；括号内为与平均值同号的随机种子数。L8 是四随机种子开发集选择；在 n=3 条件的三个随机种子上，该规则选出 L11。 *95% Singer 自助法区间不包含零（未针对 MP3 128 kbit/s 和 Opus 32 kbit/s 计算）。 †在混响条件下，A14 上无法区分 L4 和 L8，而融合头在 A14 上的最佳层是 L9。

		主要等错误率（A09–A13）				A14 等错误率（域外攻击）	
条件	n	最佳层	L10 等错误率	L19 − L10	L8 − L10	最佳层	L8 − L10
干净条件	4	L10	7.50	+1.99 (4/4)	+2.32* (4/4)	L8	−0.54 (3/4)
20分贝噪声	3	L10	15.66	+7.34* (3/3)	+2.26* (3/3)	L8	−2.18* (3/3)
10分贝噪声	3	L10	22.64	+6.63* (3/3)	+1.40 (3/3)	L8	−3.23* (3/3)
5分贝噪声	4	L10	26.99	+5.05* (4/4)	+0.31 (3/4)	L8	−3.54* (4/4)
0分贝噪声	4	L9	31.57	+4.20 (4/4)	+0.25 (3/4)	L8	−3.33* (4/4)
混响（真实房间脉冲响应）	4	L12	19.46	−0.18 (2/4)	+1.12 (3/4)	L4†	−3.12* (4/4)
MP3 128 kbit/s	3	L10	7.62	+2.67 (3/3)	+3.02 (3/3)	L8	−0.22 (2/3)
MP3 32 千比特/秒	3	L10	10.64	+2.98* (3/3)	+3.14* (3/3)	L11	+0.04 (1/3)
Opus 32 kbit/s	3	L12	9.12	+2.14 (3/3)	+2.63 (3/3)	L8	−0.35 (2/3)
Opus 12 kbit/s	3	L11	13.69	+1.81 (3/3)	+3.15* (3/3)	L11	+0.36 (2/3)

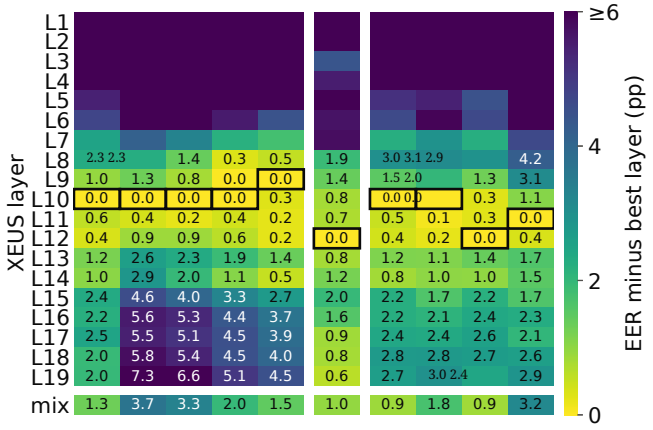

图2。仅使用 XEUS 头：在全部十种测试条件下，各层的主要等错误率减去同一条件下最佳层的主要等错误率（随机种子均值；颜色上限为 6 个百分点，显示 L8–L19 和混合层的数值）。方框标出每种条件下的最佳层。

超过 32 kbit/s 的 MP3。

混响的表现则有所不同。L10–L19 的主要等错误率均达到 18.7–20.6%，L19 − L10 为 -0.18 ± 1.44 百分点（在 4 个随机种子中，L19 有 2 个表现更好），而各随机种子对应的最佳层分别为 L19、L12、L13 和 L12。因此，上层逐渐追赶中间层，而不是使最佳层转移到新的深度。L1–L7 仍明显更差（23–29%）。在噪声条件下，学习得到的混合在每个随机种子上都劣于 L10（差距为 1.2–3.7 百分点），在两种低比特率编解码器下也同样更差（差距分别为 1.8 和 2.1 百分点）。

3.4. XEUS 的域外攻击（探索 ory)

A14 的表现有所不同。在干净音频上，其最佳 XEUS 层是 L8（13.52%），但它领先 L10 的幅度（0.54 个百分点，4 个随机种子中有 3 个）仍处于随机种子噪声范围内。在噪声条件下，L8 在 A14 上仍是最佳层，且领先幅度扩大到 2.2–3.5 个百分点；在每种信噪比下的每个随机种子上都是如此，并且在每种信噪比下重新采样歌手以及在混响条件下，这一差距依然存在（表1）。在测试片段的任一半上选择 A14 的层，结果也都是 L8；随后 L8 在另一半上胜出，留出的第四个随机种子也得出相同结论（在 5 和 0 dB 下分别为 −4.05 和 −3.26 个百分点）。相比之下，在噪声条件下，L8 的主要等错误率始终比 L10 高 0.3–2.3 个百分点。

在混响条件下，表现最佳的 A14 层向前移至 L4–L8（L4: 27.42%，L8：28.06%，L10：31.19%）。四个随机种子无法区分 L4 和 L8（$t = -0.66$），而且这种变化只出现在仅使用 XEUS 的头中；采用 CQCC 融合时，混响条件下表现最佳的 A14 层为 L9。在两种低比特率编解码器下，表现最佳的 A14 层为 L11，位于主要等错误率区间内。由于我们是在查看四个随机种子中的三个之后才选择这些比较，因此，即使对这 20 项比较进行 Bonferroni校正， p 值仍会夸大证据。我们转而依据效应量、随机种子间的一致性、拆分半组稳定性以及留出随机种子；该随机种子是在相同的测试片段上评分的。

3.5. 第二个编码器：XLS-R

图 3 按相对深度比较了两个编码器。关于 XEUS 的三项结果同样适用于 XLS-R。在所有条件下、两种随机种子下，XLS-R 的最后一层都比其最佳层差 4.4–5.5 个百分点，而且这一差距在歌手重采样后依然存在。学习得到的混合并未胜过最佳单层。在干净音频上，最佳层位于中间偏下位置（L9，相对深度为 0.38，5.71%；按随机种子分别取得最佳表现的是 L10 和 L14，其深度为 0.42 和 0.58）。有两项结果有所不同。在噪声条件下，XLS-R 没有稳定的最佳层：两个随机种子分别选择了 L7 和 L15，二者之间的等错误率几乎持平。在混响条件下，XEUS 的上层追赶上来（最后一层减最佳层：0.62 个百分点）；而 XLS-R 的最佳层在两个随机种子下都下移至 L6–L7（L7：19.57%），其上层表现变差：相对于 L7，L10 差 3.7–4.4 个百分点，L24 差 5.0–5.4 个百分点。最后，在干净音频上，XLS-R 的 A14 等错误率随深度几乎持平（L5–L21：11.5–12.5%），因此第 3.4 节的 A14 结果是 XEUS 特有的。

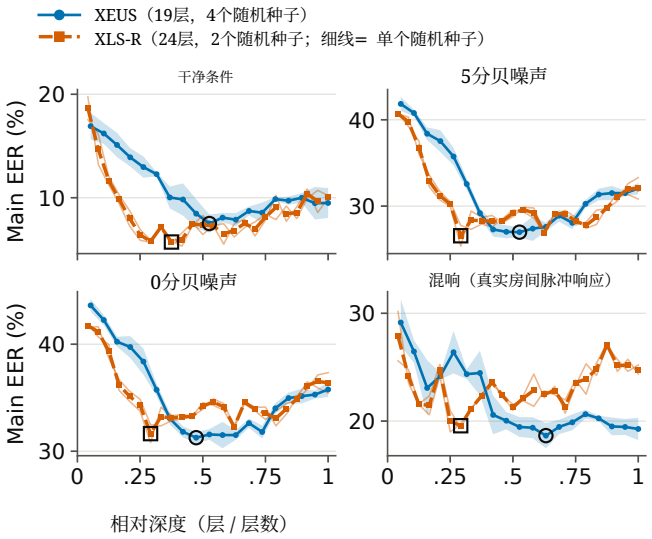

图3。主要测试等错误率随相对深度（层索引/层数）的变化，条件包括干净音频、5 dB与0 dB噪声以及混响。XEUS：随机种子均值 ± 样本标准差（4个随机种子）；XLS-R：随机种子均值（粗体）与单个随机种子（细线，2个随机种子）。空心标记：各条件下随机种子均值最优的层。

4. 讨论与局限

对于两个编码器而言，中间层都比最后一层更适合作为干净音频的读出层，但最佳层所在的位置，以及其在损坏条件下的变化，取决于编码器。采用这一方案时，在主要等错误率上，学习得到的混合并未达到最佳单层的表现。这并不表明单层总是更好：在干净条件下，XEUS 混合确实优于开发集选定层的主要等错误率，但在 A14 或噪声条件下则并非如此。

在开发集上选择层会使干净条件下的主要等错误率增加2.3个百分点（XEUS）和4.1个百分点（XLS-R）。这并不意味着开发数据毫无用处：对XEUS而言，开发集所选的L8同时也是干净条件下A14的最优层。开发集可能失效，是因为其错误率过低而无法对各层排序，或因其攻击与测试攻击不同，或两者兼有；我们的实验无法区分这些原因。在留出的攻击族或受损音频上验证或许有帮助，但我们尚未对此进行测试。

对于 XEUS，在噪声和编码条件下，L9–L12 在 A09–A13 上表现良好；在混响条件下，L10–L19 的表现大致相当，而 XLS-R 更偏好较浅的层。这与 XEUS 的去混响预训练 [11], 相符，但这些编码器在架构、训练数据和规模方面也存在差异，因此我们无法将这一差异归因于该目标。

本研究涵盖一个歌声数据集、两个冻结编码器和一种头部配置。由于 XLS-R 只有两个随机种子、四种测试条件且没有 CQCC融合，因此我们描述这两个编码器，而不对其进行排名。主等错误率欺骗样本来自 11 位歌手，因此若干主等错误率差异（包括干净条件下 XEUS L19 − L10 的差距）在进行歌手重采样后并不再成立。我们使用一个噪声语料库、一组真实房间脉冲响应和两种编解码器，训练过程中不进行数据增强。XEUS 在每种条件下有 3–4 个随机种子，而随机种子之间的差异同时混合了初始化和数据顺序的影响。A14 分析属于事后分析，其额外随机种子是在相同测试片段上进行评分的；攻击效应与源效应未加以区分。音乐预训练编码器和微调编码器仍未经过测试。

5. 结论

在 CtrSVDD 上，两个冻结语音编码器的中间层在干净音频上的歌声深度伪造检测效果优于其最后一层，而学习式层混合并未超过最佳单层。在开发集上选择层时，两个编码器都未能选出最佳测试层。最佳层对损坏的响应取决于编码器；对于 XEUS，域外攻击偏好一个不同于其他攻击的层。因此，逐层结果应跨条件和攻击类型进行报告，而不应假定最后一层或开发集错误率最低的层能够迁移。

6. 遵守伦理标准

本研究使用了依据各自许可协议公开可用的数据：CtrSVDD（CC BY-NC-ND 4.0），其真实来源语料库分别遵循各自的条款；MUSAN（CC BY 4.0）；以及 RIRS NOISES（Apache 2.0），其中的真实脉冲响应来自 RWCP、AIR 和 REVERB 2014，并遵循各自的条款。本研究不再分发任何音频，也未招募或录制新的人类参与者。

7. 致谢

本研究部分得到了 NSF ACCESS 计划分配编号 CIS261055 所提供的计算资源支持。我们感谢 Sound Ethics 的 James O’Brien 对本研究的指导。Claude（Anthropic）和 ChatGPT（OpenAI）为稿件清晰度和修订提供了支持。作者还使用这些工具审核代码，并为图1–3编写绘图代码。使用这些工具后，作者根据需要审阅并编辑了相关内容，并对出版物的内容承担全部责任。

8. 参考文献

[1] X. Wang 和 J. Yamagishi，“探究用于语音欺骗对策的自监督前端”，载于*Proc. Odyssey*，2022年。[2] H. Tak、M. Todisco、X. Wang、J.-w. Jung、J. Yamagishi 和 N. Evans，“使用 wav2vec 2.0 和数据增强进行自动说话人验证欺骗与深度伪造检测”，载于*Proc.Odyssey*，2022年。[3] B. Hurt 等，“用于跨域语音深度伪造检测的冻结SSL编码器和后端特性分析”，载于*Proc. IEEE ISM*，2026年，即将发表。[4] A. Pasad、J.-C. Chou 和 K. Livescu，“自监督语音表征模型的逐层分析”，载于*Proc. IEEE ASRU*，2021年。[5] Y. El Kheir、Y. Samih、S. Maharjan、T. Polzehl 和 S. M¨oller，“用于音频深度伪造检测的SSL模型全面逐层分析”，载于*Findings of NAACL*，2025年，第4070–4082页。[6] S.-w. Yang 等，“SUPERB：语音处理通用性能基准”，载于*Proc. Interspeech*，2021年。[7] X. Liu 等，“ASVspoof 2021：迈向真实环境中的欺骗语音与深度伪造语音检测”，*IEEE/ACM Trans. Audio, Speech,Lang. Process.*，第31卷，第2507–2522页，2023年。[8] Y. Zang、Y. Zhang、M. Heydari 和 Z. Duan，“SingFake：歌唱声音深度伪造检测”，载于*Proc. IEEEICASSP*，2024年。

[9] Y. Zang、J. Shi、Y. Zhang、R. Yamamoto、J. Han、Y. Tang、S. Xu、W. Zhao、J. Guo、T. Toda 和 Z. Duan，“CtrSVDD：受控歌唱声音深度伪造检测的基准数据集与基线分析”，载于*Proc. Interspeech*，2024年，第4783–4787页。[10] Y. Zhang、Y. Zang、J. Shi、R. Yamamoto、T. Toda 和 Z. Duan，“SVDD 2024：首届歌唱声音深度伪造检测挑战赛”，载于*Proc.IEEE SLT*，2024年。[11] W. Chen、W. Zhang、Y. Peng、X. Li、J. Tian、J. Shi、X. Chang、S. Maiti、K. Livescu 和 S. Watanabe，“迈向面向数千种语言的鲁健语音表征学习”，载于*Proc. EMNLP*，2024年，第10205–10224页。[12] A. Babu 等，“XLS-R：大规模自监督跨语言语音表征学习”，载于*Proc. Interspeech*，2022年。[13] J. Peng、O. Plchot、T. Stafylakis、L. Mo˘sner、L. Burget 和 J. ˘Cernock´y，“一种基于注意力的后端，可实现Transformer模型在说话人验证中的高效微调”，载于*Proc.IEEESLT 2022*，2023年，第555–562页。[14] Y. El Kheir、A. Das、E. E. Erdogan、F. Ritter-Guttierez、T. Polzehl 和 S. M¨oller，“两种视角，一个真相：用于鲁棒语音深度伪造检测的频谱特征与自监督特征融合”，载于*Proc. IEEE WASPAA*，2025年。[15] D. Snyder、G. Chen 和 D. Povey，“MUSAN：音乐、语音与噪声语料库”，arXiv:1510.08484，2015年。[16] T. Ko、V. Peddinti、D. Povey、M. L. Seltzer 和 S. Khudanpur，“混响语音数据增强对鲁棒语音识别的研究”，载于*Proc. IEEE ICASSP*，2017年，第5220–5224页。