arXiv:2611.32457v1 [cs.SD] 24 Nov 2026

层级排名在静音裁剪后仍然有效吗？音频深度伪造检测中冻结的 WavLM 表征里的静音捷径

努尔·埃拉比德[1] 和马赫穆德·埃萨德·阿拉尔[2]

[1]法蒂赫·苏丹·穆罕默德·瓦克夫大学生物医学工程系，土耳其伊斯坦布尔。[2]法蒂赫·苏丹·穆罕默德·瓦克夫大学电气与电子工程系，土耳其伊斯坦布尔。

通讯作者： mearar@fsm.edu.tr；贡献作者：nur.elabid@stu.fsm.edu.tr；

摘要

对冻结的自监督编码器进行逐层分析，可用于定位音频深度伪造检测中的欺骗信息。在 ASVspoof 2019 逻辑访问（LA）基准上，真实录音的前导和尾随静音比伪造语音录音更长，检测器会利用这一线索，即静音捷径。本文考察冻结的 WavLM Base 编码器的层排名在裁剪这些静音后是否仍然成立。我们在每个 12 层上分别使用裁剪和未裁剪的音频训练一个小型分类器，并在 ASVspoof 2019 LA 上进行评估，同时以零样本方式在 In-the-Wild 上评估，涵盖全部四种训练/测试组合，并计算说话人级自助法区间。在 40 分贝阈值下进行裁剪，会使每一层的域内错误率上升，且只能部分保留层排名（域内 Kendall 的 τ 为 0.58，在 In-the-Wild 上为 0.73–0.76）。40 分贝下的最佳层保持不变，但在评估集上，15 分贝时最佳层从第5层变为第6层；使用均值池化时，20 分贝阈值下也会发生这一变化。在这些阈值下，裁剪也可能移除低能量语音。在 40 分贝下，对于使用包含静音的音频训练的模型，仅裁剪其测试音频时，每一层的等错误率都会上升 6.3–19.9 个百分点。在 In-the-Wild 上，这种影响取决于层：在第3–5层，只有对使用裁剪音频训练的模型裁剪测试音频才会降低错误率；而在第6、8和9层，使用裁剪音频训练也会降低干净测试音频上的错误率。我们量化了零填充如何通过组归一化改变 WavLM 特征，而注意力掩码无法完全消除这一影响。

关键词： 音频深度伪造检测、语音欺骗对策、自监督语音表征、WavLM、逐层分析、静音捷径

1 引言

语音合成和语音转换系统 [1] 已发展到合成语音往往令听者难以与真实录音区分的程度。这项技术在无障碍、娱乐和人机交互领域有正当用途，但也助长了针对语音身份验证的冒充攻击、欺诈性音频证据以及大规模虚假信息的传播 [2]。自动说话人验证（ASV）系统尤其容易受到此类攻击，因为这些系统旨在拒绝冒充者，而非已注册说话人的合成复制品 [3]。因此，欺骗对策（CM）是实际应用中的必要措施，而ASVspoof挑战赛系列已建立标准化基准，用于评估这些对策 [4]。

目前大多数欺骗对策都采用大型自监督学习（SSL）语音模型作为前端，例如wav2vec 2.0 [5], HuBERT [6], WavLM [7], 或 XLS-R [8]„ 通常与检测头 [9, 10]联合微调。联合优化使得难以区分预训练表示已编码的信息与微调新增的信息 [11]。冻结编码器可以消除这一混杂因素：表示保持固定，一次前向传播即可获得每一层的输出，而不同配置之间的差异可归因于所测试的下游选择。因此，对冻结或部分适配的编码器进行逐层分析，已成为探究与欺骗检测相关的信息位于何处的一种常见方法 [11–15]。据报道，对于 WavLM Base，中间层的信息量最大；在近期对六种冻结的自监督学习模型进行的比较中，第4层是表现最佳的单层 [15]。

这些结论建立在存在已知捷径的基准上。在 ASVspoof 2019 逻辑访问（LA）数据库中，真实语音录音的前导和尾随静音时长比伪造语音更长，检测器会利用这一线索。M¨uller 等人 [16]表明，仅凭静音时长就能以远高于随机水平的准确度区分这两类语音，而且移除静音会使训练好的检测器错误率提高数倍。Zhang 等人 [17] 分析了这一影响在不同攻击类型中的表现；ASVspoof 2021 综述 [18]以及语音欺骗检测中捷径学习的更广泛研究 [19]也讨论了这一问题。如果某一层排名靠前，部分原因是它能很好地编码静音，那么在这一基准上测得的层排名，部分反映的就是各层读取静音的能力。两项研究与检验这一问题较为接近。Wang 等人 [20] 使用有静音和无静音的音频训练并测试了 10 个 CM，其中包括微调的 wav2vec 2.0 系统，涵盖四种训练/测试组合中的三种，但没有研究冻结编码器或单独的层。Xiao and Vu [21] 在有静音和无静音的条件下，基于冻结的 XLS-R 训练了逐层分类器，并通过余弦距离分布表明，静音会影响所有层，在 In-the-Wild 数据上对浅层的影响最大。他们仅以图表形式报告了一种训练条件下的逐层错误率，也未检验各层的排序是否保持不变。冻结编码器各层的排名能否在静音裁剪后保持不变，尚未经过检验。

本文使用冻结的 WavLM Base 研究这一问题。针对 12 个 Transformer 层中的每一层，使用 ASVspoof 2019 LA 音频的均值池化或统计特征池化表示训练双层分类器；音频要么保持原样（干净），要么移除前导和尾随静音（修剪后），每种配置使用五个随机种子。模型在 ASVspoof 2019 LA 评估集上测试，并且不进行任何适配，直接在 In-the-Wild 数据集上测试 [22]。对于每一层，使用干净音频训练的模型在干净音频上测试，使用修剪后音频训练的模型则在修剪后音频上测试。每个模型还会在另一种条件下测试，由此为每一层构成完整的 2×2 训练/测试设计。仅使用开发集选择层和检查点，并另外报告第 4 层作为参考配置。通过说话人级自助法估计不确定性。本文的目的并非降低错误率，而是说明冻结编码器逐层行为在多大程度上取决于静音。本文的贡献如下。

1.**静音裁剪的逐层影响及排名保持性测试**。我们报告了每个 WavLM Base 层在包含和不包含静音时的等错误率（EER），并定量测试 12 层的排名在裁剪后是否仍能保持。结果表明，这种保持性只是部分成立：在 40 分贝阈值下，ASVspoof 2019 LA 评估集上，干净排名与修剪后排名之间的 Kendall's τ 为 0.58，而 In-the-Wild 上为 0.73–0.76；在更严格的 20 和 15 分贝阈值下，该数值下降（评估集上为 0.36–0.48，In-the-Wild 上为 0.48–0.64）。在 40 分贝下，最佳层没有变化（评估集上为第5层，In-the-Wild 上为第4层）；而第1层和第2层在评估集上的 EER 增幅最大，排名也降至接近末尾。在 15 分贝下，以及采用均值池化的 20 分贝下，评估集上的最佳层变为第6层（在 20 分贝下采用统计池化时，第5层和第6层几乎相等，分别为 5.80% 和 5.82%），而第4层在 In-the-Wild 上仍为最佳层。

2. **分离裁剪在训练时与测试时的影响**。早期研究涵盖了完整反制措施的四种训练/测试组合中的部分或全部组合 [16, 20, 23]。我们将完整的 2×2 设计应用于冻结编码器的 12 层，并将每种效应定义为单元格之间的明确差值。在评估集上，仅裁剪测试音频时，使用干净音频训练的模型的等错误率上升 6.3–19.9 个百分点，且每一层的区间均不包含零（40 分贝）。在第3–5层，20 和 15 分贝下的增幅为 10.6–19.1 个百分点。在 In-the-Wild 上，这种效应取决于层。在第3–5层，只有对使用裁剪音频训练的模型，裁剪测试音频才会降低错误率；在三个阈值下的六种配置中，这种交互效应均具有显著性。在第6、8、9层，使用裁剪音频训练也能降低模型在干净测试音频上的错误率。在 40 分贝下，按原始时长对 In-the-Wild 进行划分后发现，对于第3–5层，这种测试时效应集中在时长不超过 4 秒分析窗口的录音中，因此，裁剪导致的窗口偏移无法解释这些层上的结果。

3. **零填充会改变 WavLM 特征**。将输入填充至固定长度会通过其卷积编码器的组归一化层，改变 WavLM Base 表示本身。在随机抽取的500个训练文件中，第4层相对于未填充特征的平均相对偏差，对于干净音频为0.31，对于

裁剪音频为0.44；传入注意力掩码可以减少这种偏差，但无法将其完全消除。因此，本研究中的所有特征均在未填充的情况下提取。

4. **逐攻击分析与说话人级不确定性**。报告了 13 种评估攻击各自的裁剪效果，并给出说话人级自助法置信区间，同时进行域外评估。在 40 分贝下，等错误率的变化幅度从 A10 的上升 16.7–24.3 个百分点，到 A17 小幅但显著的下降不等。在 20 和 15 分贝下，A19 也出现显著下降。

这些结果带来两点建议。针对 ASVspoof 2019 LA 的逐层研究应同时报告干净条件和裁剪条件，因为浅层的排名，以及在更严格的裁剪阈值下评估集上的最佳层，都会受到静音的影响（第 4.4 节和 4.7节）。提取 WavLM Base 特征时不应使用零填充，以确保文件的表示仅取决于其自身的采样点：填充会通过卷积编码器的组归一化改变特征，而注意力掩码无法完全消除这种变化。这项建议并非基于性能：在所比较的两种配置中，填充在一些测试集上降低了错误率，在另一些测试集上则提高了错误率（第 4.2节）。

本文其余部分安排如下。第2 节回顾相关工作。第3 节介绍数据、裁剪与提取流程、分类器、2×2 设计及统计分析。第4节展示结果，第5节将结果与先前工作进行比较，并说明局限性；第6 节作总结。

2 相关工作

2.1 用于欺骗检测的自监督前端

早期的欺骗对策将手工设计的倒谱特征（如恒定Q变换或线性频率倒谱系数）与高斯混合模型或轻量级卷积网络相结合 [4, 24, 25]；这些系统在已知攻击上表现良好，但面对未见过的合成算法时性能会下降。直接处理原始波形的端到端模型，尤其是基于时频图注意力的模型，在这方面取得了显著进步 [26, 27]。随后，Wang 和 Yamagishi [9] 证明，预训练自监督学习前端的表现远胜于手工设计的特征，而对前端进行微调还能进一步提升性能；此后出现的系统则将 WavLM 或 XLS-R 与注意力、敏感层或 Conformer后端相结合 [10, 28, 29]。这些系统的话语级聚合方式包括均值池化和统计池化 [30] ，以及学习得到的注意力池化 [31, 32]；本文仅采用两种固定变体。

2.2 自监督学习表征的逐层分析

对自监督语音模型进行的逐层分析表明，声学属性在早期层中最容易获取，而音素和词级信息则在中间层达到峰值 [33]。在语音欺骗检测方面，El Kheir等人 [11] 分析了多种语言及深度伪造数据集中的多个自监督学习模型，发现较低层所携带的与欺骗检测相关的信息多于最终层。Pan等人 [12]将 WavLM Large 的隐藏嵌入与注意力机制结合，并报告称其较低层至中间层对决策的贡献最大；Sinha等人 [13]提出了

层条件嵌入融合；Spoof-SUPERB 基准比较了冻结的自监督学习模型与通过学习得到的跨层加权和 [34]；Beheshti 等人 [14]则在特定任务训练前使用探测分类器对层进行排名，并报告称信息丰富的层聚集在不同深度区域。其他研究探讨了这些表示的时间动态，而非其逐层结构 [35]。近期有两项研究直接考察了冻结的 WavLM Base。Serrano等人 [15] 比较了六种冻结的自监督学习编码器在域外数据上的层选择和融合方法，发现第4层是 WavLM Base 的最佳单层。Mallat 等人 [36] 使用 ASVspoof 5，对 WavLM Base 各层经均值和标准差池化后的输出训练线性分类器，发现第11层作为单层时表现最佳，并通过投影加权典型相关分析将各层与17种声学属性关联起来。这些研究均未比较有无静音时得到的逐层结果，尽管 Beheshti 等人在训练前会裁剪静音。

2.3 静音作为捷径

M¨uller 等人 [16] 指出，ASVspoof 2019 LA 中的真实语音话语，其前导和尾随静音时长显著长于伪造语音；仅使用前导静音时长的模型，其表现远高于随机水平；而修剪静音会显著改变训练后的检测器的错误率。Zhang 等人 [17]将这一效应归因于静音的比例和内容两方面，并指出文本转语音（TTS）系统生成的语音中静音比例较低；他们还发现，即使重新训练检测器，移除静音后端到端文本转语音攻击仍更难检测。ASVspoof 2021 挑战赛包含一些隐藏评估子集，其中的非语音区间经语音活动检测移除；排名靠前的系统在这些子集上的性能有所下降，且对文本转语音攻击的下降幅度大于对语音转换攻击的下降幅度 [18]。Shim 等人 [19]提出了一种通用框架，在训练数据和测试数据中引入受控干预，以揭示对策中的捷径学习；Rubio 等人 [37]则为 XLS-R 检测器开发了一种基于干预的诊断方法，其中包括对非语音片段和零填充的操纵。为在静音被修剪时保持鲁棒性而设计的方法包括结合多任务元学习的数据增强 [20] 和多子空间注意力 [38]。这些研究将对策视为一个整体，并未考察各个编码器层。Xiao 等人 [39] 将冻结的 XLS-R 和 HuBERT 编码器神经元激活中伪造语音与真实语音的差异投影到静音与语音子空间上，发现该子空间在 ASVspoof 2019 评估集上可解释 13–14% 的差异（按有效秩计），而在 In-the-Wild 上仅能解释 0.3–0.5%；他们的结果汇总了所有层，且未比较训练/测试条件。

2.4 最相关的工作

与本研究最接近的研究总结于表1中。Wang 等人 [20]在三种训练/测试组合下评估了十种对策，包括微调的wav2vec 2.0系统：在未修改音频上训练和测试、在未修改音频上训练而在裁剪音频上测试，以及在裁剪音频上训练和测试，并采用

表1 采用有静音和无静音数据训练并测试对策的研究。各单元格表示干净音频和裁剪音频的训练/测试组合

	M¨uller [16]	Wang [20]	Shim [23]	Xiao [21]	Ours
编码器	CQT 对策模型，RawNet2	10 个对策模型	AASIST-L	XLS-R	WavLM Base
冻结编码器	—	—	—	✓	✓
逐层等错误率	—	—	—	仅图	✓
训练/测试单元格	[a]中的3	4个中的3个	4个中的4个	4 个中的 2 个	4 个中的 4 个（所有层）
层排序测试	—	—	—	—	✓
In-the-Wild	—	—	—	✓	✓

[a]不同模型之间；对于 RawNet2，未说明测试音频是否经过修剪。

15 分贝的修剪阈值。他们还仅使用静音训练了一个模型，并分析了每次攻击的静音时长；此外，他们在ASVspoof 2021的不同数据集上测试了自己的方法。他们的研究不包括冻结编码器、逐层分析，也没有在In-the-Wild数据上进行评估。Xiao and Vu [21] 为冻结的XLS-R编码器的每一层分别连接了一个分类器，融合了逐层决策，并比较了有静音和无静音时的训练效果。他们利用余弦距离的分布表明，静音会影响所有层，并且在In-the-Wild数据上对浅层的影响最大。他们的训练包含数据增强，但没有报告两种训练条件下的逐层错误率，没有采用完整的训练/测试设计，也没有测试层排序是否发生变化。M¨uller 等人 [16] 评估了一个使用静音训练、在裁剪后测试音频上进行测试的模型，并在裁剪音频上训练了RawNet2；Shim 等人 [23]报告了AASIST-L的全部四种训练/测试组合，使用一种基于能量的检测器移除超过50毫秒的静音；上述研究均未逐层分析冻结编码器。本研究在全部四种训练/测试组合下，使用固定分类器且不采用数据增强，测量冻结编码器每一层的等错误率，并定量检验层排序是否保持不变（表1）。

3 方法

图 1 总结了该处理流程。波形会根据需要修剪前导和尾随静音，在不进行填充的情况下限制为最长 4 秒，然后输入冻结的 WavLM Base 编码器。单个层的隐藏状态会在时间维度上进行池化，并由一个小型前馈网络进行分类；该网络是唯一经过训练的组件。

3.1 数据

分类器在 ASVspoof 的逻辑访问分区上训练 2019 [4]，并在其评估分区以及未经任何适配的 In-the-Wild 数据集上测试 [22]（表2）。使用 ASVspoof 2019 LA，是因为该数据库中的静音线索已有文献记录和分析 [16, 17],，因此可以将结果与这些研究直接比较。ASVspoof 2019 LA 的各分区按说话人划分，彼此不重叠。训练集和开发集包含六种攻击（A01–A06）；评估分区包含十三种攻击（A07–A19），其中只有 A16 和 A19 重用了训练中出现过的算法（分别为 A04 和 A06 的算法）。In-the-Wild 包含从公开来源收集的公众人物真实语音和伪造语音录音，其信道和录音条件不受控，而且

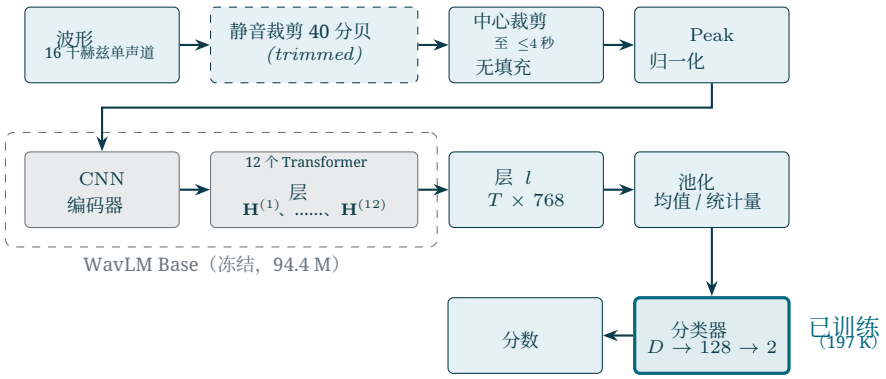

图1 检测流程。仅在裁剪条件下应用静音裁剪（虚线）。每个文件均以自身长度输入编码器，最长为 4 秒，不进行零填充。编码器被冻结；仅训练分类器

表2 数据集。训练集、开发集和评估分区来自 ASVspoof 2019 LA；In-the-Wild 仅用于零样本测试

Set	真实语音	伪造语音	总计	攻击	Use
训练集	2,580	22,800	25,380	A01–A06	分类器训练
开发集	2,548	22,296	24,844	A01–A06	检查点和层选择
评估	7,355	63,882	71,237	A07–A19	域内测试
In-the-Wild	19,963	11,816	31,779	—	域外测试

从不用于训练或任何选择。说话人身份是说话人级自助法（第3.8节）所必需的信息，其来源是 ASVspoof 2019 的官方协议文件，以及 In-the-Wild 的数据集元数据。评估分区中，有 67 位说话人提供了真实语音文件，48 位说话人提供了伪造语音文件；In-the-Wild 的每个类别均包含 54 位说话人。

3.2 预处理和静音裁剪

每个文件都以 16 kHz 的单声道波形加载。我们比较两种条件。在干净条件下，波形保持原样使用。在修剪后条件下，先使用`librosa.effects.trim`（librosa 0.11.0）裁剪前导和尾随静音：该函数会丢弃文件开头和结尾处能量比最响帧低 40 分贝以上的帧（帧长为 400 个采样点，帧移为 160 个采样点）。话语内部的停顿不受影响。如果裁剪后剩余时长不足 0.5 秒，则保留未裁剪的文件；训练集有 8 个、开发集有 5 个、评估集有 21 个、In-the-Wild 文件有 49 个符合此情况。按照 M¨uller et al. [16] 的做法，本文所报告实验的阈值在实验开始前固定；在 40 分贝阈值下，任何集合中回退到未裁剪文件的比例至多为 0.15%。随后，两种条件均执行相同的两个步骤。时长超过 4 秒的文件裁剪为居中的 4 秒（64,000 个采样点）；较短的文件保留实际长度，不进行填充。短于卷积编码器感受野（400 个采样点）的文件会填充至 400 个采样点；训练集、开发集或 In-the-Wild 集合中均没有此类文件

。最后，每个波形都按峰值归一化为$\mathbf{x}/(\max_n |x_n| + 10^{-8})$。

为考察结果对阈值的依赖程度，我们使用 20 和 15 分贝的阈值重复了裁剪条件，并采用相同的帧设置、回退规则，以及相同的特征提取、训练和分析流程。这两个阈值是在获得 40 分贝的结果以及测量 40 分贝裁剪后剩余的静音量之后添加的（第 4.7 节）；因此，本文将其作为敏感性分析报告，所有其他结果均对应 40 分贝。阈值为 20 分贝时，训练集、开发集、评估集和 In-the-Wild 中分别有 80、54、158 和 708 个文件回退至未裁剪文件；阈值为 15 分贝时，这些数字分别为 132、105、279 和 1,147。也就是说，回退文件最多占 ASVspoof 2019 LA 文件的 0.52%，以及 In-the-Wild 文件的 3.6%。在 In-the-Wild 中，真实语音文件的回退频率高于伪造语音文件（20 分贝时分别为 2.83% 和 1.22%，15 分贝时分别为 4.70% 和 1.76%）；在 ASVspoof 2019 LA 中，两种类别的文件回退比例为 0.21–0.54%。回退文件会保留静音，因此在更严格的阈值下，In-the-Wild 中保留静音的真实语音文件多于伪造语音文件。在任一阈值下，均无文件短于 400 个采样点。

3.3 冻结特征提取

使用 WavLM Base [7]提取特征。采用公开发布的检查点 microsoft/wavlm-base；该模型由一个卷积编码器和随后连接的 12 个 Transformer 层组成，隐藏维度为 768，总计约 9440 万个参数；模型在 960 小时的 LibriSpeech 数据上进行了预训练。所有编码器参数均被冻结，并在评估模式下运行编码器（禁用随机失活、LayerDrop 和 SpecAug-ment），且不计算梯度。对于包含 T 帧的输入（每 20 毫秒一帧），编码器会为每个 Transformer 层 $l \in \{1, \ldots, 12\}$生成隐藏状态 $\mathbf{H}^{(l)} \in \mathbb{R}^{T \times 768}$；不使用卷积编码器的输出。每种条件下只提取一次特征并进行缓存，因此后续每个实验只训练分类器。

每个文件都单独通过编码器处理（批次大小为 1），且不使用注意力掩码，因此绝不会添加零填充。对于 WavLM Base，这一选择很重要。其卷积编码器的第一层会对每个通道应用组归一化，组数与通道数相同（512），因此每个通道的均值和方差都是在输入的所有时间步上计算的，其中包括填充的零。相比之下，注意力掩码只在 Transformer 层内部起作用，不会作用于卷积编码器。因此，对文件进行填充也会改变其未填充部分的表示。两种标准的批量输入处理方式都会填入零：发布的检查点在预处理器配置中要求提供注意力掩码（返回注意力掩码为 true），同时指定使用组归一化（特征提取归一化为 `"group"`）[1]，；而 Transformers 文档指出，采用组归一化的模型在训练时没有使用注意力掩码，输入只需用零填充[2]。该行为在此模型的软件实现中已有记录[3]，Rubio 等人 [37] 也在诊断 XLS-R 检测器时纳入了对零填充的操作。

[1] https://huggingface.co/microsoft/wavlm-base [2] https://huggingface.co/docs/transformers/model_doc/wav2vec2 [3] https://github.com/pytorch/audio/issues/2242

第 4.2 节逐层测量具有组归一化的编码器在特征层面的影响。填满 4 秒窗口、因此无需填充的文件，在填充提取时接收的输入与此完全相同，所以该流程只有在处理较短文件时才有所不同。特征提取使用 PyTorch 2.11.0 和 Hugging Face Transformers 5.17.0，在 NVIDIA T4 GPU 上进行。由于发现计算环境的不同会话之间的 Transformers 版本发生了变化，因此固定了该版本；5.17.0 和 5.18.0 生成的特征完全相同。

3.4 池化与分类器

所选层的帧序列 $\mathbf{h}_1, \ldots, \mathbf{h}_T$ 通过两种固定池化操作之一，缩减为定长向量。均值池化的计算方式为

$$\boldsymbol{\mu} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_t \in \mathbb{R}^{768}, \tag{1}$$

统计池化 [30] 会附加逐元素标准差，

$$\boldsymbol{\sigma} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (\mathbf{h}_t - \boldsymbol{\mu})^2}, \qquad \mathbf{v}_{\text{stats}} = [\boldsymbol{\mu}; \boldsymbol{\sigma}] \in \mathbb{R}^{1536}. \tag{2}$$

对于每个数据集，标准差均采用总体定义（除以 T）。池化后的向量由一个两层网络进行分类，

$$\text{Linear}(D, 128) \to \text{ReLU} \to \text{Dropout}(0.3) \to \text{Linear}(128, 2), \tag{3}$$

其中，均值池化使用 $D = 768$，统计池化使用 $D = 1536$，分别得到 98,690 和 196,994 个可训练参数。由于训练集中伪造文件数量约为真实文件的九倍，交叉熵损失按 $w_c = N/(2N_c)$ 加权，其中 N 为训练文件总数，N_c 为类别 c 中的文件数（真实语音约为 4.92，伪造语音约为 0.56）。检测分数是伪造类别的 Softmax概率，因此分数越高表示伪造语音；在官方评估脚本中，其符号会被反转，因为这些脚本要求真实试验具有更高的分数。

3.5 训练与模型选择

每个分类器均使用 Adam 进行训练（学习率为 10^{-4}，权重衰减为 10^{-5}），批次大小为 128，共训练 50 个训练轮次，并在这 50 个训练轮次中对学习率进行余弦退火。分别在第 1 、 5 、 10 、 ..... 、 50 个训练轮次后计算开发集等错误率（共评估 11 次），并保留数值最低的检查点。在此选择过程中，等错误率采用基于 ROC 的实现进行计算，该实现仅使用 ROC 曲线的角点；而本文报告的每个等错误率均使用官方 ASVspoof 2019 评估脚本计算。每种配置（层、池化、条件）均使用五个随机种子（42–46）进行训练，共计 $12 \times 2 \times 2 \times 5 = 240$ 次训练运行。

报告了两类配置的结果。开发集选择的 配置针对每种条件和池化方法，使用在随机种子上取平均后开发集等错误率最低的层。 参考 配置包括第 4 层采用统计池化的干净条件，以及第 4 层采用均值池化的裁剪条件。对于域外数据，Serrano 等人报告冻结的 WavLM Base 在单层上的最佳结果为第 4 层。 [15], 该层也用于此处理流程的早期版本，这些版本使用零填充。在那些版本中，干净配置的层是在评估集上选择的，而裁剪配置则是在 In-the-Wild 等错误率最低的配置。由于这两种选择都涉及测试数据，因此报告参考配置仅用于保持连续性；任何结论都不单独依赖于这些配置。依赖于所选配置的结论均基于开发集选择的配置。在此处报告的实验中，不会在评估集或 In-the-Wild 上选择训练轮次、层或配置。

3.6 训练/测试设计

令 $E(a, b)$ 表示模型在条件 a 下训练、在条件 b 下测试时的等错误率，其中 $a,b \in \{\mathrm{C}, \mathrm{T}\}$ 分别代表干净和裁剪后。对于全部 12 层和两种池化方法，均计算全部四个单元格，共得到 24 种配置；在第4.7节所述的更严格阈值下，跨条件单元格仅针对第 3–5 层计算，这些层包括参考层（第 4 层）以及在开发集上选出的层（第 3 层和第 5 层）。检查点始终根据训练条件下的开发集选取。两个跨条件单元格 $E(\mathrm{C}, \mathrm{T})$ 和 $E(\mathrm{T},\mathrm{C})$ 需要额外进行 240 次训练运行（24 种配置、两个单元格、五个随机种子）；这些模型还对各自训练条件下的测试数据进行了评分，且所得分数与对应网格模型的分数完全相同。裁剪的影响定义如下

$$\Delta_{\mathrm{test}} = E(\mathrm{C}, \mathrm{T}) - E(\mathrm{C}, \mathrm{C}), \tag{4}$$

$$\Delta_{\mathrm{train}} = E(\mathrm{T}, \mathrm{T}) - E(\mathrm{C}, \mathrm{T}), \tag{5}$$

$$\Delta_{\mathrm{total}} = E(\mathrm{T}, \mathrm{T}) - E(\mathrm{C}, \mathrm{C}), \tag{6}$$

$$\Delta^{\mathrm{C}}_{\mathrm{train}} = E(\mathrm{T}, \mathrm{C}) - E(\mathrm{C}, \mathrm{C}), \tag{7}$$

$$\Delta^{\mathrm{T}}_{\mathrm{test}} = E(\mathrm{T}, \mathrm{T}) - E(\mathrm{T}, \mathrm{C}), \tag{8}$$

$$I = \Delta^{\mathrm{T}}_{\mathrm{test}} - \Delta_{\mathrm{test}} = \Delta_{\mathrm{train}} - \Delta^{\mathrm{C}}_{\mathrm{train}}. \tag{9}$$

Δ_{test} 表示仅裁剪在干净音频上训练的模型的测试音频所产生的影响；Δ_{train} 表示在测试音频已裁剪的情况下，同时裁剪训练音频所产生的影响；$\Delta_{\mathrm{total}} = \Delta_{\mathrm{test}} + \Delta_{\mathrm{train}}$ 表示同时裁剪训练音频和测试音频所产生的影响； $\Delta^{\mathrm{C}}_{\mathrm{train}}$ 表示仅裁剪训练音频、而测试音频保持干净所产生的影响； $\Delta^{\mathrm{T}}_{\mathrm{test}}$ 表示裁剪在裁剪音频上训练的模型的测试音频所产生的影响；交互效应 I 则衡量测试时裁剪的影响在多大程度上取决于训练条件。上述六项均以等错误率的百分点表示，并针对上述 24 种配置计算。

由于分析窗口是每个文件的中间4秒，因此将时长超过4秒的文件进行裁剪，主要是移动窗口，而不是从其中移除静音。为区分这两种影响，进一步将

In-the-Wild按原始时长划分为时长不超过4秒的文件和时长超过4秒的文件，并针对每个子集计算第3–5层的四个单元格。

3.7 评估指标

主要指标为等错误率（EER），即真实试验的漏检率等于伪造语音试验的虚警率时所对应的工作点，使用官方 ASVspoof 2019 脚本计算。每个随机种子分别计算 EER，然后取平均值；结果报告为五个随机种子的均值 $\pm$ 样本标准差。

在 ASVspoof 2019 评估集上，还通过最小归一化串联检测代价函数（t-DCF），评估该反欺骗措施与自动说话人验证系统配合使用时的表现；此处采用 ASVspoof 2019 [4, 40]所使用的公式。计算时使用官方 ASVspoof 2019 脚本以及随数据库提供的官方自动说话人验证分数。后来又引入了修订后的公式 [3]；本文采用 2019 年的公式，以便结果与已发表的 ASVspoof 2019 结果保持可比。对于反欺骗措施的阈值 s，

$$\text{t-DCF}(s) = \frac{C_1\, P_{\text{miss}}^{\text{cm}}(s) + C_2\, P_{\text{fa}}^{\text{cm}}(s)}{\min(C_1, C_2)}, \tag{10}$$

其中 $C_1 = \pi_{\text{tar}}\big(C_{\text{miss}}^{\text{cm}} - C_{\text{miss}}^{\text{asv}} P_{\text{自动说话人验证漏检率}}\big) - \pi_{\text{non}} C_{\text{fa}}^{\text{asv}} P_{\text{自动说话人验证误接受率}}$ 以及 $C_2 = C_{\text{fa}}^{\text{cm}}\, \pi_{\text{spoof}}\big(1 - P_{\text{自动说话人验证漏检率，伪造语音}}\big)$。自动说话人验证错误率 $P_{\text{自动说话人验证漏检率}}$、$P_{\text{自动说话人验证误接受率}}$ 以及 $P_{\text{自动说话人验证漏检率，伪造语音}}$ 均固定在自动说话人验证系统的等错误率阈值处，最小值则在 s 上取值。采用 ASVspoof 2019 代价模型：$\pi_{\text{spoof}} = 0.05$、$\pi_{\text{tar}} = 0.9405$、$\pi_{\text{non}} = 0.0095$、$C_{\text{miss}}^{\text{asv}} = C_{\text{miss}}^{\text{cm}} = 1$ 和$C_{\text{fa}}^{\text{asv}} = C_{\text{fa}}^{\text{cm}} = 10$。与等错误率一样，每个随机种子都会计算最小 t-DCF，并报告均值 $\pm$ 样本标准差。In-the-Wild 未提供自动说话人验证分数，因此仅报告其等错误率。

3.8 统计分析

不同随机种子之间的差异仅反映训练过程中的随机性；对于冻结编码器和简单分类器，这种随机性较小。测试文件和说话人的选择所带来的不确定性，通过两个层级的自助法进行估计，每个层级重采样 1,000 次，且均按类别分层。在文件层级，每个类别内有放回地抽取文件。在说话人级，每个类别内有放回地抽取说话人，并纳入其所有文件；该层级考虑了同一说话人所说文件之间的相似性，因此得到的区间更宽，也更符合实际。在每次重采样中，均按计算点估计时的相同方式，计算每个随机种子的等错误率（EER），再对各随机种子取平均；95% 区间取自第 2.5 和第 97.5 百分位数。两个系统或单元格之间的差异在相同的重采样样本上计算（配对自助法）；若其区间不包含零，则称差异具有统计显著性。

通过 Kendall 的 τ 和 Spearman 的 ρ，利用每层的种子平均 EER 及说话人级自助法区间，衡量在干净到干净和修剪后到修剪后单元格中得到的层级排名之间的一致性（覆盖 12 层）；在每次重采样中，还会记录最佳层。对于 τ 和 ρ，百分位区间

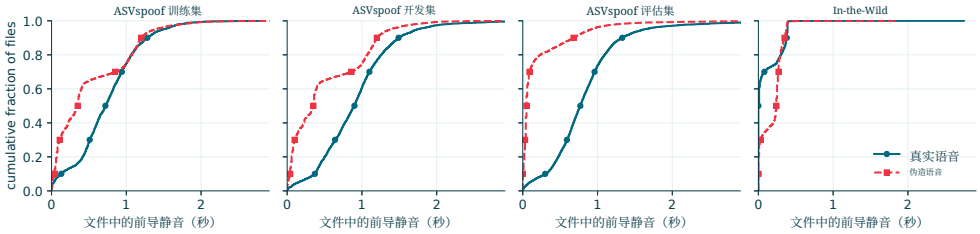

图 2 ASVspoof 2019 LA 训练集、开发集和评估集以及 In-the-Wild 中，真实语音（青绿色、实线、圆形标记）和伪造语音（红色、虚线、方形标记）文件的整段前导静音经验累积分布（40 dB 阈值）

并不以点估计为中心：在 40 分贝下，69–84% 的重采样结果给出小于 τ观测值的结果（并列时计为一半）。根据相同重采样结果计算的偏差校正加速法（BCa）区间 [41] 见表 B1；其中，加速量通过在每个类别内依次留出一名说话人进行估计。在 40 分贝下，它们得出相同结论，但其偏差校正常数在表 B1的十二种配置中介于 -0.46至 $+1.21$ 之间；有 6–23% 的重采样结果等于观测到的 τ，而该值在 12 层上只能取 67 个可能值；在 40 分贝下，BCa 上限位于重采样结果的第 99.8 百分位数处或更高。因此，对于 τ 和 ρ，以及所有其他量，均报告百分位区间。对于 p的精确双侧 τ 值，也给出了 $n = 12$；该检验将 12 个层级 EER 视为固定值，检验其两种排序是否相关，而不对文件或说话人进行重采样。与其他情况一样，每个评估攻击的攻击级 EER 均根据该攻击的伪造文件和评估集中的全部真实文件，按随机种子分别计算，再对五个随机种子取平均；裁剪所造成的变化则使用文件级和说话人级的相同配对类别分层自助法进行检验。

4 结果

4.1 数据中的静音线索

ASVspoof 2019 LA 报告的静音线索 [16] 存在于此处使用的文件中。在用于裁剪的 40 dB 阈值下，评估集上整个文件的前导静音中位数，真实语音为 0.77 秒，伪造文件为 0.05 秒；在 4 秒分析窗口内则分别为 0.72 秒和 0.04 秒；两种分布明显分离（图2）。为量化这一线索，我们仅使用静音特征，在训练分区上训练了一个逻辑回归分类器：整个文件及 4 秒窗口内的前导和尾随静音（静音），并可选择加入样本数和填充量（静音+时长）。特征经过对数变换（$\log(1+x)$）和标准化，类别权重按其频率的倒数设置，且未对音频本身进行裁剪。未使用任何音频模型时，该分类器在评估集上以 30 分贝测量静音时达到 18.82% 的等错误率，以 40 分贝测量时达到 23.40%（表3）。在 In-the-Wild 上，相同分类器的等错误率介于 49.20% 和 58.48 % 之间；该数据集中，真实语音的前导静音中位数为 0.00 秒，伪造文件为 0.24 秒。

表3 在 ASVspoof 2019 LA 训练集上训练的逻辑回归分类器的等错误率（%）：仅使用静音特征（sil：整个文件以及4秒窗口内的前导和尾随静音），或使用这些特征加上样本数和填充量（sil+dur）

特征	阈值	开发集	评估	In-the-Wild
sil	30 分贝	31.90	18.82	49.20
sil	40 分贝	30.22	23.40	58.39
静音+dur	30 分贝	31.12	18.75	49.24
静音+dur	40 分贝	34.50	23.25	58.48

4.2 零填充对 WavLM 特征的影响

为测量第3.3节所述的影响，我们从每种条件下随机抽取的 500 个训练文件以及填充量最大的 500 个训练文件中提取特征，采用三种方式：不填充（参考）、填充至 4 秒且不使用注意力掩码，以及填充至 4 秒并使用注意力掩码。相对于参考值的偏差，定义为平均绝对差除以参考特征的平均绝对值。图3 展示了每一层的结果。在随机抽取的文件中，第4层的填充使干净音频的特征变化量达到其自身幅度的 0.31，裁剪音频则达到 0.44。注意力掩码将这些数值降至 0.15 和 0.34，但并未消除这一影响。裁剪音频受到的影响更大，因为裁剪会缩短文件并增加填充量（中位数为 1.76 秒，而干净音频为 0.85 秒）；且影响随填充量增加而增大：对于第4层未使用掩码的裁剪音频，填充量少于 1 秒时影响约为 0.20，超过 3 秒时约为 0.94。

填充本身并不是一种强捷径。仅使用填充量（该量取决于文件时长）的分类器，在两种条件下的训练集、开发集、评估集和 In-the-Wild 集合上，得到的等错误率均为 41.6–48.9%。不使用填充进行特征提取也不会降低错误率。对于两个参考配置，带填充的特征提取（在干净条件下不使用注意力掩码，在裁剪条件下使用注意力掩码）得到的评估等错误率分别为 1.23%（clean/stats/第4层）和 6.42%（trimmed/mean/第4层）；不使用填充时则分别为 1.48% 和 6.44%。其真实场景等错误率分别为 18.92% 和 14.00%，而不使用填充时分别为 18.26% 和 15.12%（五个随机种子的均值）。两种处理流程使用相同的分类器、训练设置、随机种子、检查点选择和官方等错误率脚本；它们的差异仅在于特征提取方式。因此，不使用填充进行特征提取并非出于性能考虑，而是因为这样得到的文件表示只取决于该文件自身的采样点，而不取决于分析窗口的长度，也不取决于批次级填充情况下批次中的其他文件。

4.3 逐层结果与 2×2 设计

图 4 展示了两个匹配单元中每一层的等错误率。在评估集上，裁剪条件在每一层的等错误率都高于干净条件，并且差距在前几层最大（统计池化，第1层：干净条件为3.15%，裁剪条件为15.97%）。在In-the-Wild上，趋势则相反：裁剪会降低大多数层的等错误率，尤其是第2–9层，而第11层和第12层基本保持不变。

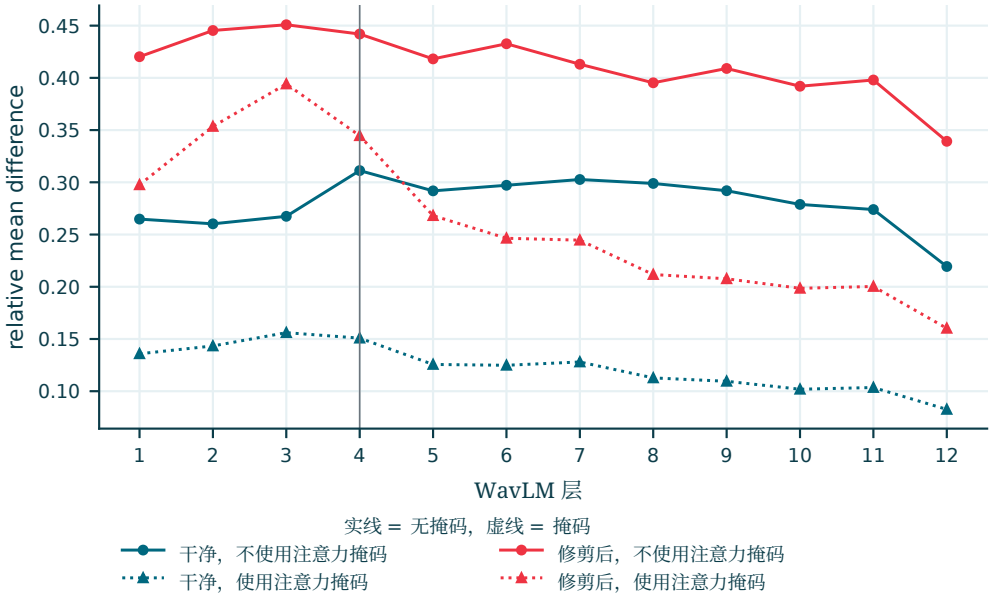

图 3 在干净和裁剪音频上，对 500 个随机抽取的训练文件，比较每层填充与未填充 WavLM Base 特征的相对偏差（青绿色：干净；红色：裁剪），分别为不使用（实线）和使用（点线）注意力掩码。竖线标示第4层

均值池化与统计池化之间的差异始终很小。在干净条件下以及In-the-Wild的两种条件下，表现最佳的三层均为第3–5层；而在评估集的裁剪条件下，则为第4–6层；第4层在条件与池化方式的四种组合中，In-the-Wild等错误率均为最低。

表4 列出了由开发集选定的配置和参考配置。开发集在干净条件下选出第3层，在裁剪条件下选出第5层。六种配置中，裁剪均会使评估集上的等错误率上升 3.7–7.2 个百分点，且每种情况下，在文件级和说话人级上均达到显著性。评估集上的最低串联 t-DCF 也呈现相同变化趋势：干净条件下为 0.0416–0.0528，裁剪条件下则为 0.1290–0.1667。在 In-the-Wild 上，裁剪使等错误率下降 2.0–4.4 个百分点；六种配置在文件级上的下降均显著，但在说话人级上，仅统计特征/第3层、均值/第5层和统计特征/第5层达到显著性。In-the-Wild 上每个类别有 54 名说话人，说话人级区间的跨度约为 $\pm$5个百分点。

图6 和表5 给出了 Δ_{test} 以及全部 12 层的交互效应，图5展示了第3–5层的四个单元格。在评估集上，仅从测试音频中修剪静音时，使用干净音频训练的模型的等错误率上升 6.27–19.88 个百分点（ Δ_{test} ）；在全部 24 种配置中，区间均不包含零；使用两种池化方法时，第4层的增幅最小，第1层的增幅最大。使用裁剪音频训练的模型，在测试音频也经过修剪时变化要小得多（$+0.11$ 至 $+4.13$ 个百分点；区间在 24 种配置中有 12 种不包含零，而在第5、7、9、11和12层中，两种池化方法均不包含零；表A2）。使用.....进行训练

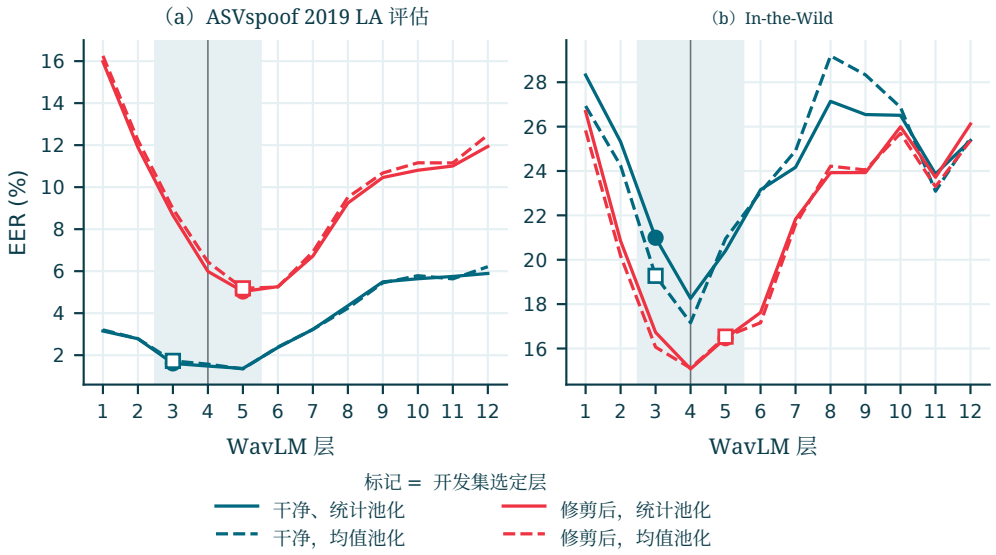

图4 ASVspoof 2019 LA 评估集 (a) 和 In-the-Wild (b) 上各 WavLM Base 层的等错误率，涵盖干净和修剪后条件（五个随机种子的均值）。青绿色：干净；红色：修剪后。实线：统计池化；虚线：均值池化。阴影带标示第3–5层，竖线标示第4层。标记表示在开发集上选定的层（实心：统计池化；空心：均值池化）

表4 开发集选定配置和参考配置：EER（%）及最小 tandem t-DCF，以五个随机种子的均值 ± 样本标准差表示；EER 附带 95% 说话人级自助法区间。选择：开发集 = 开发集选定，参考 = 参考配置

		Dev	ASVspoof 2019 LA 评估			In-the-Wild	
配置	Sel.	EER	EER	95%置信区间	最小化归一化反欺骗对策	EER	95%置信区间
干净/均值/L3	开发集	0.44 ± 0.02	1.73 ± 0.01	[1.36, 2.11]	0.0528 ± 0.0002	19.28 ± 0.28	[14.34, 24.47]
干净/统计量/L3	开发集	0.43 ± 0.00	1.61 ± 0.01	[1.27, 1.95]	0.0478 ± 0.0007	20.99 ± 0.38	[15.60, 26.97]
干净/统计量/第4层	ref	0.54 ± 0.01	1.48 ± 0.02	[1.15, 1.85]	0.0416 ± 0.0003	18.26 ± 0.48	[13.64, 23.27]
修剪后/均值/第4层	参考	1.89 ± 0.01	6.44 ± 0.01	[5.47, 7.47]	0.1667 ± 0.0007	15.12 ± 0.18	[10.80, 19.99]
修剪后/均值/第5层	开发集	1.60 ± 0.02	5.18 ± 0.02	[4.52, 5.92]	0.1390 ± 0.0008	16.54 ± 0.07	[12.45, 21.15]
修剪后/统计数据/第5层	开发集	1.60 ± 0.02	5.03 ± 0.04	[4.29, 5.75]	0.1290 ± 0.0006	16.46 ± 0.14	[12.57, 21.06]

使用裁剪音频训练、使用干净音频测试时，等错误率上升 2.03–8.91 个百分点，且所有 24 种配置的区间均不包含零。对于统计量/第4层，四个单元格分别为 $E(\mathrm{C},\ \mathrm{C}) = 1.48$ %、 $E(\mathrm{C},\ \mathrm{T}) = 7.75\%$、 $E(\mathrm{T},\ \mathrm{C}) = 5.37\%$ 和 $E(\mathrm{T},\ \mathrm{T}) = 5.99\%$。交互效应介于 -5.65 和 -15.74 个百分点之间，且为负值；在全部24种配置中，区间均不包含零。因此，使用干净音频训练的模型在每一层都受静音影响，而使用裁剪音频训练的模型受影响要小得多。

在 In-the-Wild 上，对使用干净音频训练的模型的测试音频进行裁剪，在24种配置中的任何一种配置下，EER的变化最多为1.11个百分点；其中三种配置的说话人级区间不包含零，且符号不一。对训练音频和测试音频同时进行裁剪

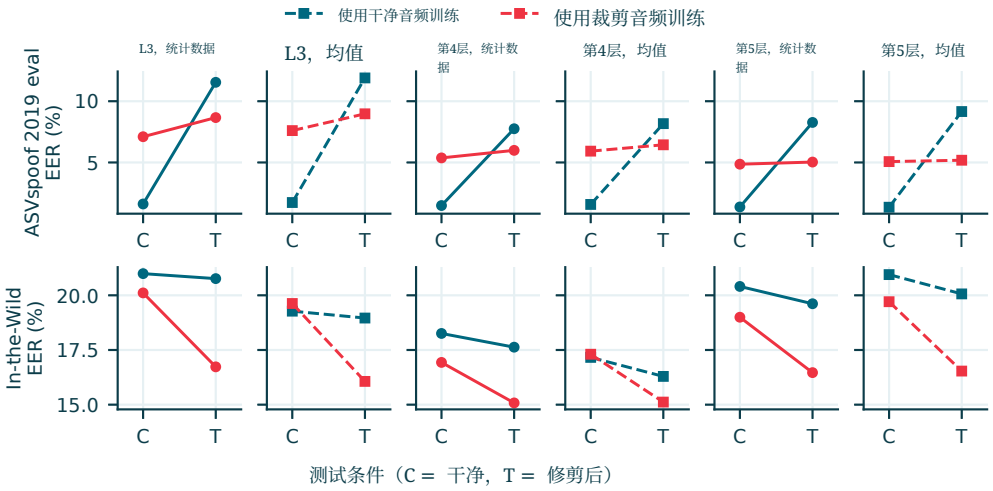

图 5 第3–5层在四种训练/测试单元格中的等错误率（EER）。这些层包括开发集选出的第3层和第5层，以及参考层（第4层）；结果采用均值池化和统计池化，并对五个随机种子取均值。所有层的效果见图 6 和表 5。横轴为测试条件；每条线代表一种训练条件（青绿色：使用干净音频训练；红色：使用修剪后音频训练）。实线：统计池化；虚线：均值池化

会使11种配置下的EER降低，且区间不包含零（第5–9层的均值池化、第3层和第5–9层的统计池化）。在另外11种配置中，交互效应为负且区间不包含零（第1–5层、第7层和第9层的均值池化，第3–5层和第9层的统计池化）；其中包括第3–5层的全部六种配置，这些层包含开发集选定的第3层和第5层，以及参考层第4层。在这六种配置中，对使用裁剪音频训练的模型的测试音频进行裁剪，会降低其EER（$\Delta^{\mathrm{T}}_{\mathrm{test}}$，降低1.85–3.56个百分点），但不会降低使用干净音频训练的模型的EER。在第6、8和9层，即使测试音频是干净的，使用裁剪音频训练也会降低EER（$\Delta^{\mathrm{C}}_{\mathrm{train}}$，公式（7））；两种池化方法的降幅均为2.53–5.41个百分点，且区间不包含零；在第7层，区间包含零（表A2）。对于第12层的统计池化，交互效应为正（$+1.55$ $[+0.02, +3.14]$）；在其余12种配置中，交互效应的区间包含零，其中包括第6、8、10和11层的所有配置。在第12层，对使用裁剪音频训练的模型的测试音频进行裁剪会使EER升高，均值池化和统计池化分别升高0.73和0.91个百分点；第8层的统计池化也有同样情况（升高0.95个百分点），且区间不包含零（表A2）。这些区间未针对24次比较进行调整。表5 报告了统计池化的结果；均值池化的数值见附录A，其中还给出了第3–5层的文件级和说话人级区间（图A1）。对于统计量/第4层，各单元格分别为$E(\mathrm{C},\mathrm{C}) = 18.26\%$、$E(\mathrm{C},\mathrm{T}) = 17.63\%$、$E(\mathrm{T},\mathrm{C}) = 16.93\%$和$E(\mathrm{T},\mathrm{T}) = 15.08\%$。

4.4 层排名的保持

表6 比较了干净到干净和修剪后到修剪后的单元格中 12 层的排名。在评估集上，排名仅部分保留（$\tau = 0.58$）。在 In-the-Wild 上，τ 为 0.73–0.76，但其区间较宽（表6），因此一致性

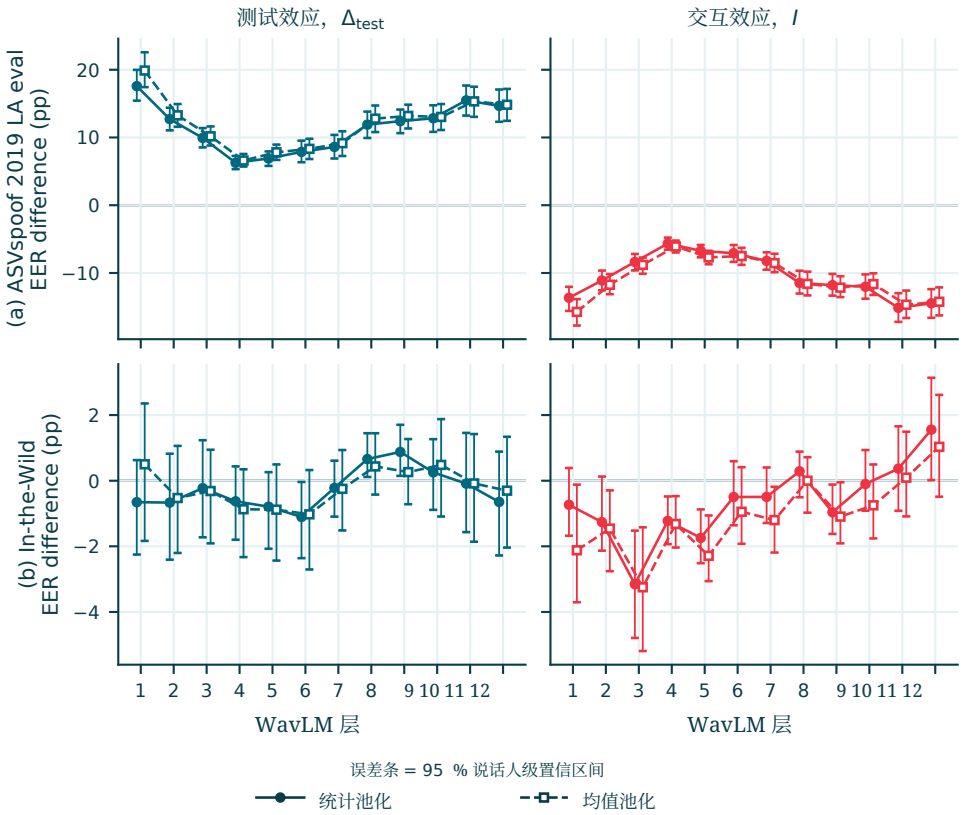

图 6 在 ASVspoof 2019 LA 评估集（a，顶行）和 In-the-Wild（b，底行）上按层进行裁剪的影响：对使用干净音频训练的模型裁剪测试音频（Δ_{test}，公式（4）；左列）以及交互效应 I（公式（9）；右列）。统计池化：实线、实心圆；均值池化：虚线、空心方块，并略微向右偏移。误差条表示 95% 的说话人级配对自助法区间；每行共用纵轴。各随机种子的等错误率，取五个随机种子的均值

没有直接比较两个测试集上的结果。最佳层没有变化：评估集上为第5层，In-the-Wild 上为第4层，在两种条件和两种池化方法下均如此。自助法重采样中该层排名最佳的比例介于 55%（修剪后、均值池化、评估集）和 99% 之间。在 15 分贝下，以及在采用均值池化的 20 分贝下，评估集上的最佳层为第6层（第4.7节）。

差异源于最初几层（表 7）。在采用统计池化的评估集上，层1和层2在裁剪后等错误率增幅最大（分别为12.8和9.1个百分点），排名也从第6和第5降至第12和第10。层7–12的等错误率大致翻倍（$\times$1.9–2.1），相对顺序保持不变。

表5 裁剪的影响（等错误率百分点，统计池化），附 95% 说话人级配对自助法区间。不包含零的区间以粗体显示。评估：ASVspoof 2019 LA 评估集；ITW：In-the-Wild。定义见公式（4）–（9）

Set	L	Δ_{test}	Δ_{train}	Δ_{total}	I
评估	1	**+17.58 [15.44, 19.98]**	**−4.76 [−6.67, −3.20]**	**+12.82 [10.97, 14.58]**	**−13.66 [−15.60, −12.04]**
	2	**+12.71 [11.05, 14.39]**	**−3.60 [−5.08, −2.11]**	**+9.12 [7.96, 10.38]**	**−11.11 [−12.56, −9.64]**
	3	**+9.94 [8.53, 11.39]**	**−2.88 [−4.09, −1.75]**	**+7.06 [5.98, 8.18]**	**−8.37 [−9.62, −7.19]**
	4	**+6.27 [5.33, 7.28]**	**−1.76 [−2.56, −0.95]**	**+4.51 [3.73, 5.38]**	**−5.65 [−6.55, −4.78]**
	5	**+6.90 [5.80, 7.94]**	**−3.24 [−4.10, −2.35]**	**+3.67 [3.06, 4.26]**	**−6.73 [−7.68, −5.86]**
	6	**+7.87 [6.35, 9.53]**	**−4.98 [−6.19, −3.80]**	**+2.90 [2.18, 3.62]**	**−7.08 [−8.36, −5.87]**
	7	**+8.60 [6.90, 10.38]**	**−5.11 [−6.41, −3.93]**	**+3.49 [2.67, 4.42]**	**−8.21 [−9.51, −6.95]**
	8	**+11.90 [9.93, 13.82]**	**−7.00 [−8.48, −5.50]**	**+4.89 [3.85, 5.89]**	**−11.47 [−13.04, −9.67]**
	9	**+12.41 [10.64, 14.12]**	**−7.44 [−8.74, −6.04]**	**+4.97 [3.89, 6.01]**	**−11.79 [−13.34, −10.13]**
	10	**+12.83 [10.83, 14.75]**	**−7.67 [−9.06, −6.07]**	**+5.17 [3.99, 6.26]**	**−12.03 [−13.80, −10.21]**
	11	**+15.47 [13.23, 17.67]**	**−10.22 [−12.08, −8.30]**	**+5.26 [4.05, 6.45]**	**−15.14 [−17.22, −12.98]**
	12	**+14.67 [12.32, 17.08]**	**−8.62 [−10.74, −6.47]**	**+6.05 [4.81, 7.46]**	**−14.47 [−16.61, −12.39]**
ITW	1	−0.65 [−2.25, 0.63]	−0.99 [−6.90, 5.92]	−1.64 [−7.37, 5.08]	−0.73 [−1.68, 0.39]
	2	−0.67 [−2.41, 0.82]	−3.80 [−8.28, 1.28]	−4.47 [−8.30, 0.05]	−1.26 [−2.13, 0.13]
	3	−0.23 [−1.73, 1.23]	−4.03 [−7.61, 0.10]	**−4.26 [−8.01, −0.53]**	**−3.15 [−4.79, −1.52]**
	4	−0.63 [−1.80, 0.44]	−2.55 [−5.89, 0.92]	−3.18 [−6.53, 0.29]	**−1.23 [−1.93, −0.48]**
	5	−0.79 [−2.07, 0.26]	**−3.15 [−5.64, −0.46]**	**−3.94 [−6.54, −1.60]**	**−1.74 [−2.51, −0.87]**
	6	**−1.11 [−2.36, −0.04]**	**−4.43 [−6.53, −1.72]**	**−5.53 [−7.28, −3.28]**	−0.50 [−1.36, 0.59]
	7	−0.22 [−1.09, 0.61]	−2.12 [−4.68, 0.59]	**−2.34 [−4.71, −0.17]**	−0.49 [−1.29, 0.40]
	8	**+0.66 [0.11, 1.45]**	**−3.87 [−6.71, −0.86]**	**−3.21 [−6.07, −0.15]**	+0.29 [−0.50, 0.89]
	9	**+0.88 [0.15, 1.70]**	**−3.49 [−5.58, −1.40]**	**−2.61 [−4.80, −0.54]**	**−0.96 [−1.62, −0.12]**
	10	+0.25 [−0.89, 1.27]	−0.77 [−3.48, 2.21]	−0.52 [−3.13, 2.06]	−0.10 [−0.91, 0.94]
	11	−0.09 [−1.57, 1.46]	−0.06 [−2.97, 3.30]	−0.15 [−2.77, 2.60]	+0.37 [−0.92, 1.66]
	12	−0.65 [−2.28, 0.89]	+1.39 [−1.72, 5.38]	+0.74 [−2.04, 4.08]	**+1.55 [0.02, 3.14]**

表6 干净到干净单元格与修剪到修剪单元格之间的层级排名一致性，以及95%的说话人级自助法区间。给出了最佳层及其为最佳层的重采样比例

Set	池化	Kendall 的 τ	p （精确）	Spearman 的 ρ	最佳层（干净 / 修剪后）
评估	统计数据	0.58 [0.39, 0.64]	0.009	0.66 [0.52, 0.76]	第5层（82%）/ 第5层（82%）
评估	mean	0.58 [0.42, 0.64]	0.009	0.66 [0.51, 0.76]	第5层 (99%) / 第5层 (55%)
In-the-Wild	统计数据	0.73 [0.42、0.85]	0.0005	0.90 [0.57, 0.94]	第4层 (90%) / 第4层 (82%)
In-the-Wild	mean	0.76 [0.39, 0.82]	0.0002	0.87 [0.51、0.94]	第4层（91%）/ 第4层（69%）

4.5 按原始时长划分的 In-the-Wild

表 8 将 In-the-Wild 划分为不超过4秒的文件（19,387个）和超过4秒的文件（12,392个）。对于时长超过窗口的文件，裁剪主要只是移动窗口，而仅裁剪测试音频不会产生影响（$\Delta_{\text{test}} = -0.08 \pm 0.09$）。对于能够放入窗口的文件，裁剪会使等错误率降低 0.90 ± 0.08 个百分点（五个随机种子的均值和标准差），但这一效应的说话人级区间包含零；在此处以及采用两种池化方法的第3层和第5层中，情况均如此。对于使用裁剪音频训练的模型，裁剪测试音频会降低短文件的等错误率，并且在所有这些配置中，说话人级区间均不包含零。测试时效应是

表 7 评估集上各层的等错误率（%），统计池化，在干净到干净和修剪后到修剪后单元格中的结果（五个随机种子的均值）

层	干净	修剪后	比率	排名（干净 →修剪后）
1	3.15	15.97	×5.1	$6 \to 12$
2	2.78	11.90	×4.3	$5 \to 10$
3	1.61	8.67	×5.4	$3 \to 5$
4	1.48	5.99	×4.0	$2 \to 3$
5	1.37	5.03	×3.7	$1 \to 1$
6	2.37	5.26	×2.2	$4 \to 2$
7–12	3.23–5.89	6.72–11.94	.9–2.1	相同相对顺序

表 8 In-the-Wild 按原始时长划分，统计池化，第4层。每个单元格的等错误率（%）及其影响（百分点）（均值 ± 五个随机种子上的样本标准差）

	≤4 秒（19,387 个文件）	>4s（12,392 个文件）
E（C，C）	20.34	16.65
E（C，T）	19.44	16.57
E（T，C）	19.65	12.81
E（T，T）	17.43	12.70
Δ_{test}	-0.90 ± 0.08	-0.08 ± 0.09
Δ_{train}	-2.01 ± 0.30	-3.88 ± 0.26
Δ_{total}	-2.91 ± 0.26	-3.95 ± 0.27

因此，效果主要集中在短文件中；对于基于裁剪音频训练的模型，这一现象尤为明显，而且窗口的移动无法解释这些结果。两个子集在实现改进所需的条件上也有所不同。对于长文件，仅使用裁剪音频进行训练就足够了： E(T，C) $=$ 12.81%，而 E(C，C) $=$ 16.65%。对于短文件，训练和测试时都需要进行裁剪才能实现改进。第 3 层和第 5 层以及均值池化也呈现相同的模式。

4.6 逐攻击分析

裁剪对不同攻击的影响并不相同（图 7，表 9）。每种攻击贡献了来自48名说话人的4,914个伪造文件，并与评估集中的所有真实文件进行评分。A10受影响最大，在六种配置下分数增加16.7–24.3个百分点，等错误率从约3–4.5%升至20–29%。其次是A11和A15，分别增加4.1–14.4和4.5–10.9个百分点。A07、A08、A09、A12、A14和A16增加0.6–6.3个百分点，且在全部六种配置下均达到显著水平。其余变化较小：A13增加+0.04 至 +0.12 个百分点（除统计量/第4层外均显著）；A18增加 +0.2 至 +1.0（统计量/第5层下不显著）；A19增加 +0.1 至 +0.5 个百分点（仅在

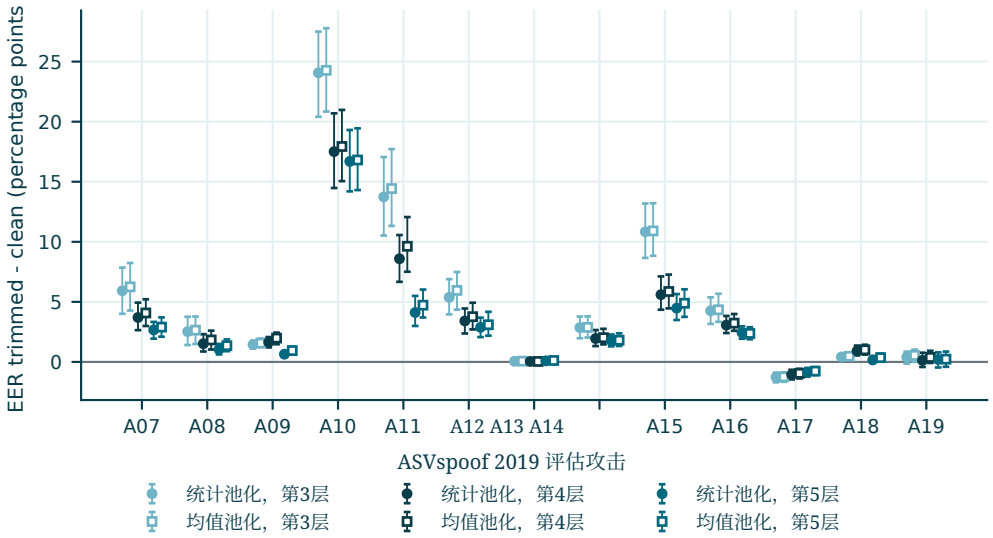

图7 裁剪后各评估攻击及六种配置的等错误率变化（修剪后减去干净），以及 95% 说话人级自助法区间。颜色表示层（天蓝绿色：L3；深青绿色：第4层；青绿色：第5层）；实心圆表示统计池化；空心方块表示均值池化

表9 各攻击裁剪后的等错误率变化（修剪后减去干净，百分点）；六种配置的范围

组	攻击	变化
受影响最大	A10	+16.7 至 +24.3
大	A11, A15	+4.1 至 +14.4； +4.5 至 +10.9
中等	A07, A08, A09, A12, A14, A16	+0.6 到 +6.3
小	A13；A18；A19	+0.04 到 +0.12； +0.2 到 +1.0； +0.1 到 +0.5
下降	A17	−0.76 至 −1.27

均值/第3层下显著）。A17是唯一一种裁剪后等错误率下降的攻击，降幅为0.76–1.27个百分点，并且在全部六种配置下，两个层级的下降均显著。对大多数攻击而言，从第3层到第5层，增幅随层数加深而减小；例如，采用统计池化时，A11的增幅分别为13.7、8.6和4.1个百分点，A15则分别为10.8、5.6和4.5个百分点。例外是A09和A18，它们在第4层的增幅最大；A13和A19的变化则非常小。因此，这些攻击与真实语音之间的区分最依赖干净条件下的静音。

4.7 对修剪阈值的敏感性

在 40 分贝处进行裁剪会削弱静音线索，但不会将其完全消除。在裁剪音频上测量时，当以 15 分贝阈值定义静音时，评估集中保留的前导静音中位数为：真实语音 0.16 秒，伪造文件 0.03 秒，

而在 20 分贝下分别为 0.14 秒和 0.02 秒。第 4.1 节的仅静音分类器基于这些残差测量进行训练，在评估集上，当静音分别以 15、20 和 30 分贝（静音）进行测量时，等错误率达到 32.67%、33.48% 和 35.58%；在静音+持续时间方面，等错误率为 31.55%–34.79%，相比之下，未裁剪音频上的等错误率为 18.82% 和 23.40%（表 3）。在 In-the-Wild 上，相同分类器的等错误率介于 33.40% 和 45.62% 之间。在 40 分贝处裁剪后，每个集合中仍有 3.1%–4.2% 的文件保留超过 50 毫秒的前导静音（在 4秒窗口内为 2.1%–4.0%），且真实语音文件中的比例高于伪造文件（评估集上分别为 6.7% 和 3.0%），中位数为 0.16–0.23 秒。当一个略高于阈值的短促声音先于停顿出现时，便会产生这种残留：裁剪点位于该声音处，而其后的停顿则被保留。五个真实语音 In-the-Wild 文件的音量过低，无法在 40 分贝处进行裁剪（最响帧的均方根低于 10^{-3}，因此 librosa 的底噪下限 10^{-5} 比其低不足 40 分贝）。

更严格的阈值会从两类文件中移除更多音频。对于真实语音文件，从评估文件中移除的时长中位数从 40 分贝下的 1.17 秒增加到 20 分贝下的 1.56 秒和 15 分贝下的 1.64 秒；对于伪造语音文件，则从 0.29 秒增加到 0.45 秒和 0.51 秒；在 15 分贝下，这相当于真实语音文件时长中位数的 54%。在 In-the-Wild 数据集上，40 分贝规则从真实语音文件中移除的时长中位数为 0.01 秒，从伪造语音文件中移除 0.35 秒；而在 20 分贝下，更严格的阈值分别移除 0.71 秒和 0.63 秒，在 15 分贝下分别移除 0.78 秒和 0.71 秒。在更严格的阈值下额外移除的音频，始于 40 分贝规则检测到声音的位置，因此不一定是静音。

表10 在 20 和 15 分贝阈值下重复了主要分析。在裁剪条件下，三个阈值中，开发集选定层对于两种池化方法均为第5层；在干净条件下，则为第3层（表 4）。采用统计池化时，在每一层和每个阈值下，裁剪后的等错误率均高于干净音频上的等错误率；评估集上开发集选定配置的等错误率从 40 分贝时的 5.03% 上升到 20 分贝时的 5.80% 和 15 分贝时的 6.20%。在更严格的阈值下，条件交叉单元格仅针对第3–5层计算，因为这些层包含开发集选定层和参考层。对于这些层，评估集上的测试时效应从 40 分贝时的 6.3–10.2 个百分点增至 20 分贝时的 10.6–18.0 个百分点，以及 15 分贝时的 11.2–19.1 个百分点；统计池化在第4层的交互效应从 -5.65 到 -10.80 和 -11.15个百分点。层级排名在更严格的阈值下的一致性较低（Kendall 的 τ 精确p-值见表B1；在 20 和 15 分贝下，四种评估集配置中有三种的精确 p 超过 0.05，因此对于这三种配置，精确检验无法区分这种一致性与两个互不相关的排序之间的一致性，而百分位区间则不包含零）；在 15 分贝下，两种池化方法的评估集最佳层均从第5层变为第6层（分别在 63% 和 89% 的重采样中），在 20 分贝下采用均值池化时也从第 5 层变为第 6 层（ 95% ）；在 20 分贝下采用统计池化时，第5层和第6层几乎相同（5.80% 和 5.82%），且第6层在 56% 的重采样中表现最佳。在 In-the-Wild 上，最佳层仍为第4层，在 98%–99% 的重采样中如此。

在 In-the-Wild 上，交互效应仍为负值；在两个更严格的阈值下，第3–5层的六种配置中，说话人级区间均不包含零。然而，仅裁剪测试音频如今会使基于干净音频训练的模型的等错误率在 20 分贝下上升 0.3–2.5 个百分点、在 15 分贝下上升 1.0–3.7 个百分点；而在 40 分贝下，这些层的等错误率则降低 0.2–0.9 个百分点。除第5层的均值池化（20 和 15 分贝）以及第3层的均值池化（15 分贝）外，该效应的说话人级区间均包含零。

表10 三个修剪阈值下的主要结果。效应以百分点表示，并附有95%的说话人级自助法区间（统计池化，第4层）；Kendall's τ 比较干净和修剪后的层级排序。回退文件是指因剩余时长不足0.5秒而保持未修剪的文件。移除时长中位数是通过修剪从文件中移除的时长中位数，其中回退文件按零计。最佳层及其在重采样中被选为最佳层的比例（统计池化）一并给出。跨条件单元格以及由此产生的 Δ_{test} 和 I，仅在20和15分贝下适用于第3–5层

	40 分贝	20 分贝	15 分贝
回退文件，评估 / ITW	21 / 49	158 / 708	279 / 1,147
移除时长中位数（秒），评估，真实语音 / 伪造语音	1.17 / 0.29	1.56 / 0.45	1.64 / 0.51
移除时长中位数（秒），ITW，真实语音 / 伪造语音	0.01 / 0.35	0.71 / 0.63	0.78 / 0.71
开发集选定层，修剪后	L5	L5	L5
评估等错误率，统计数据，第5层（%）	5.03	5.80	6.20
ITW等错误率，统计数据，第5层（%）	16.46	16.42	16.67
Kendall's τ，评估，统计数据 / 均值	0.58 / 0.58	0.48 / 0.39	0.42 / 0.36
Kendall's τ，ITW，统计数据 / 均值	0.73 / 0.76	0.52 / 0.58	0.64 / 0.48
最佳层，评估	第5层（82%）	第5层（44%）	第6层（63%）
最佳层，ITW	第4层（82%）	第4层（99%）	第4层（98%）
Δ_{test}，评估，第3至第5层，两种池化方式	6.3 至 10.2	10.6 至 18.0	11.2 至 19.1
Δ_{total}，评估	4.51 [3.73, 5.38]	6.59 [5.57, 7.69]	6.79 [5.68, 7.98]
交互效应，评估	−5.65 [−6.55, −4.78]	−10.80 [−12.35, −9.32]	−11.15 [−12.70, −9.66]
Δ_{total}，ITW	−3.18 [−6.53, 0.29]	−4.55 [−8.15, −0.80]	−4.06 [−7.93, 0.002]
交互效应，ITW	−1.23 [−1.93, −0.48]	−3.83 [−5.24, −2.23]	−3.47 [−4.95, −1.80]

攻击的排序基本保持不变。A10 仍是受裁剪影响最大的攻击，在 20 分贝下增加 19.7–32.2 个百分点，在 15 分贝下增加 20.5–32.2 个百分点。经过裁剪后，在三个阈值下 A17 均能被更好地检测出来。A19 在 40 分贝下裁剪后其等错误率略有上升，但在 20 和 15 分贝下，六种配置中裁剪后均能更好地检测出 A19，且说话人级区间不包含零。

4.8 层编码的内容

我们使用来自 20 位说话人的 5,000 个训练文件（按标签和攻击类型分层抽样），基于每层的均值池化特征（768 维）训练线性探针，以预测声学属性和静音属性。声学目标——平均基频、平均谐噪比（HNR）和谱倾斜——通过 Praat 并借助 Parselmouth 库计算。岭回归采用按说话人分组的五折交叉验证进行评估，而基于打乱的目标训练的探针所得结果 R^2 接近于零。图8 和表11 总结了干净音频上的结果。声学属性从第1层预测效果最佳，随着层数加深，预测效果逐渐变差，这与此前对自监督学习模型进行的逐层分析一致 [33]。每一层对前导静音的预测效果都同样出色（$R^2 \approx 0.74$），因此每一层都包含静音线索。4秒窗口内信号的长度在每一层也都能得到良好预测：干净音频上的 R^2 介于 0.90 和 0.96 之间，裁剪音频上则至少为 0.88；然而，基于填充量（即窗口长度减去信号长度）的分类器，其表现接近随机水平（第4.2节）。这些探针仅用于描述：最擅长编码声学属性的层（第1层）并非最擅长检测语音欺骗的层（第3–5层），因此无法解释层排名。Mallat 等人 [36], 分析了冻结的 WavLM

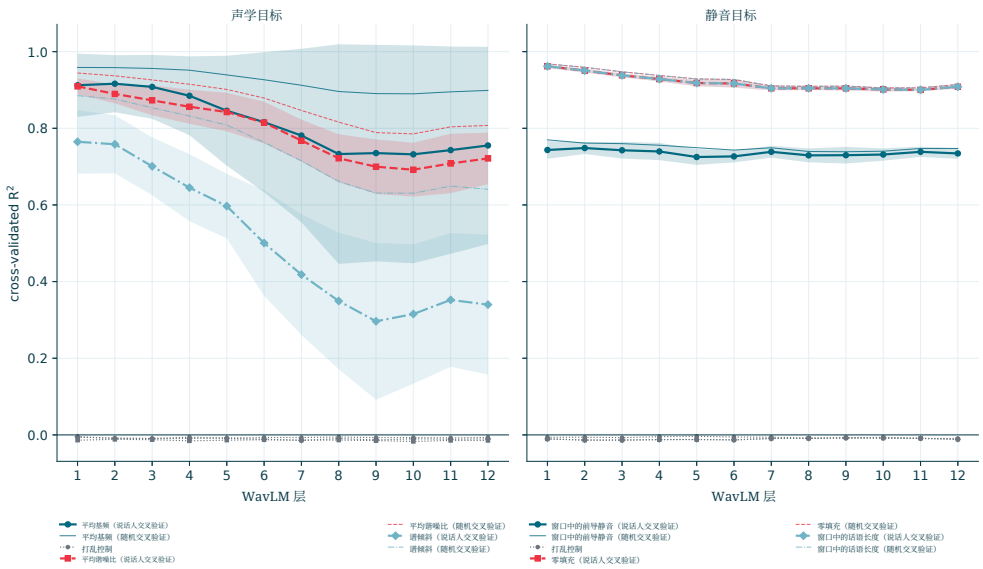

图8 干净音频上逐层的探针 R^2 。左：声学目标；右：静音和长度目标。曲线显示均值池化特征上岭回归探针的交叉验证均值 R^2 ；带标记的粗线：按说话人分组的五折交叉验证，阴影带表示五个折上的 ± 1 标准差；细线：随机五折交叉验证；灰色点线：打乱目标控制。由于填充等于窗口长度减去长度，右图中的填充和长度目标具有相同的 R^2

表11 选定层上的探针 R^2 （干净音频，按说话人分组的交叉验证）

目标	L1	L4	L8	L12
平均基频	0.91	0.88	0.73	0.76
平均谐噪比	0.91	0.86	0.72	0.72
谱倾斜	0.76	0.65	0.35	0.34
前导静音	0.74	0.74	0.73	0.73
4秒窗口中的长度	0.96	0.93	0.90	0.91

采用不同分析方法对使用 ASVspoof 5 训练的 Base 进行了研究，发现大多数声学属性在第7层附近达到峰值。

5 讨论

5.1 与既有静音研究的关系

捷径的规模与早期报告一致。M¨uller 等人 [16] 仅根据前导静音的时长，在 ASVspoof 2019 评估集上获得了 15.12% 的等错误率；训练时去除静音后，RawNet2 的等错误率则从 3.61% 升至 15.50%，约为原来的四倍。他们的描述并未明确说明评估音频是否也经过了裁剪。

仅使用静音特征，在 30 分贝下的等错误率为 18.82%（表 3）。采用统计池化时，若测试音频也经过裁剪，使用裁剪音频训练会使第3–5层的等错误率增至原来的 3.7–5.4 倍（表 7）；若测试音频干净，则增至原来的 3.6–4.4 倍（表 7 和 A2），因此无论采用哪种理解，比较结果都成立。裁剪影响最大的攻击类型也一致。Zhang 等人 [17]同样发现，A10 和 A11 的性能下降最大，并将其归因于 Tacotron 2 声学模型，因为该模型的输出缺少前导和尾随静音。ASVspoof 2021 的隐藏子集显示，文本转语音受到的性能下降大于语音转换攻击 [18],；后者在 ASVspoof 2019LA 中为 A17–A19。本研究中，A10、A11 和 A15 受影响最大；A18 和 A19 在 40 分贝下变化不大，而 A17 的性能有所提升（第4.6节）；在 20 和 15 分贝下，A19 的性能也有所提升（第4.7节）。

在 In-the-Wild 上，线索的方向则相反：在 40 分贝下，真实语音的前导静音中位数为 0.00 秒，而伪造文件为 0.24 秒（第 4.1 节，图 2），并且在 ASVspoof 2019 LA 上训练的仅使用静音的分类器，在该数据集上的等错误率介于 49.20% 和 58.48% 之间（表 3），接近或差于随机水平。

这些 2×2 结果将 Wang 等人 [20] 对 AASIST-L 的三单元格比较，以及 Shim 等人 [23] 的四单元格比较，扩展到了冻结编码器的每一层。在裁剪后音频上训练、在干净音频上测试这一单元格，将训练条件的影响与测试条件的影响区分开来。在评估集上，这表明未使用静音进行训练的模型在测试时利用静音所获得的提升很小（$\Delta^{\mathrm{T}}_{\mathrm{test}}$ 为 $+0.11$ 至 $+4.13$ 个百分点），而使用静音进行训练的模型在每一层都依赖静音。在 In-the-Wild 上，结果取决于层数：在第3–5层，仅在训练或测试时使用裁剪后音频，均不会在说话人级别上显著改变等错误率，而交互效应在六种配置中均达到显著；在第6、8和9层，仅使用裁剪后音频进行训练会显著降低等错误率。Xiao and Vu [21], 关于静音对冻结的 XLS-R 的浅层影响最大这一分层观察，与此处 WavLM Base 第1层和第2层的表现一致：裁剪后，这两层的等错误率增幅最大，排名降幅也最大。

5.2 层选择

在所有条件下，最佳层均位于第3层至第6层之间；在条件与池化的四种组合中，第4层的真实场景等错误率均最低。这与 Serrano 等人 [15], 的研究结果一致：他们发现，在使用不同训练语料（ASVspoof 5）、输入长度（3秒片段）和池化方式（均值）的情况下，冻结的WavLM Base在域外数据上的最佳单层是第4层。Mallat 等人 [36], 同样使用冻结的 WavLM Base，但在ASVspoof 5上进行训练，并为每一层训练一个线性分类器；他们发现第11层作为单层时表现最佳。这两项研究在训练数据和分类器方面存在差异，而当前结果无法确定究竟是哪一项差异导致了这一现象。探测结果表明，声学属性在第1层中最易获取，这与 Pasad等人 [33], 的研究结果一致，因此它们无法解释第3–5层的优势。

由于最佳层位于编码器的下半部分，因此推理时无需计算其上方的层。El Kheir等人 [11] 表明，只保留自监督学习前端的较低层（基础模型保留4–6层，大模型保留10–12层），所得等错误率与完整模型相当，同时参数量至少减半。

5.3 域外性能的背景

开发集选定配置的真实场景等错误率介于16.46%和20.99%之间。对于冻结编码器，Serrano 等人 [15] 报告了在 ASVspoof5 上训练的 WavLM Base 第4层取得17.8%的等错误率；Spoof-SUPERB 基准 [34] 报告了 WavLM Large 和 HuBERT Base 分别取得24.33%和27.28%的等错误率。这两个模型均在 ASVspoof 2019 LA 上训练，各自仅以固定随机种子运行一次，而此处的数值是五个随机种子的均值。采用更大的编码器和不同的训练方式，已有研究报告了显著更低的错误率：Beheshti 等人 [14] 使用 XLS-R-300M 的四个选定层、数据增强，并在训练前进行静音裁剪，取得了4.94%的等错误率（同一研究中，WavLM Large 单次运行的结果为9.68%）；Pascu 等人 [42] 使用冻结的 wav2vec2/xls-r-2b 表示和逻辑回归，取得了7.2%的等错误率。本研究的目标并非降低错误率，而是说明静音捷径如何影响冻结编码器的每一层。选择这些配置，是为了让这一分析更易于解释，而非最大限度地提升域外性能。具体而言，编码器保持冻结，分类器则保持精简，以便将不同配置之间的差异归因于层和训练条件。

5.4 影响

由此可得出两点实践启示。首先，在 40 分贝阈值下，对于包含静音的 ASVspoof 2019 LA，逐层分析得出的结论对最佳层和最深的层稳定，但对最浅的层则不然，其排名很大程度上取决于静音；在更严格的阈值下，四种配置中有三种在评估集上的最佳层也会变化，变为相邻层。报告这两种条件可以清楚地呈现这一点。其次，任何将 WavLM Base 输入填充至固定长度的处理流程，都会通过卷积编码器的组归一化改变特征，而且传入注意力掩码也无法避免这一问题；按每个文件自身的长度提取特征可以规避该问题，但代价是无法进行批处理。本研究仅测量了 WavLM Base 的这一影响，其卷积编码器使用组归一化；其他编码器是否会受到影响，取决于其采用的归一化方式，本研究未对此进行测试。

5.5 局限性

本研究存在若干局限。研究仅考察了一种编码器 WavLM Base，若使用更大的模型或其他自监督学习编码器，结果可能有所不同。In-the-Wild 中每个类别仅有 54 位说话人，因此该数据库上的说话人级区间较宽，约为 ± 5 个百分点。主要分析采用单一修剪阈值（40 分贝）；在已知 40 分贝结果后，才加入更严格的 20 和 15 分贝阈值（后者与 Wang 等人使用的阈值相同），并将其作为敏感性分析报告。40 分贝下的裁剪仍保留了部分静音线索（第 4.7节），而在 20 和 15 分贝下裁剪后剩余的静音并未测量；在这两个阈值下额外移除的音频，从 40 分贝规则检测到声音的位置开始（第 4.7节），其中可能包含话语开头和结尾处较弱的语音声响或呼吸声，这些内容未经检查。因此，在这些阈值下，裁剪条件可能移除的不只是静音；此外，这些阈值下的跨条件单元格仅针对第 3–5 层计算。研究采用了单一分析窗口（4 秒）。研究仅移除了前导和尾随静音

，这是该数据库文档中所记录的线索形式 [16]；研究未考察话语内部的停顿。分类器是一个基于冻结编码器的小型网络，经过微调的系统可能表现不同。所有模型仅使用 ASVspoof 2019LA 进行训练。

6 结论

本文考察了 ASVspoof 2019 LA 中的静音捷径如何影响冻结的 WavLM Base 编码器的逐层表现。修剪前导和尾随静音会提高每一层的域内等错误率，并且只能部分保留 12 层的排名。在 40 分贝时，最佳层保持不变；但在更严格的阈值下，评估集上的最佳层在四种配置中的三种发生了变化，而在裁剪条件下根据开发集选出的层在三个阈值下均为第5层（干净条件下为第3层）。完整的 2×2 设计表明，在域内，经过含静音数据训练的模型在每一层都会依赖静音；而在 In-the-Wild 上，裁剪的影响取决于层。零填充会通过组归一化改变 WavLM Base 特征，因此所有特征提取均未使用填充。

由此提出两点建议。针对 ASVspoof 2019 LA 的逐层研究应同时报告干净条件和裁剪条件，因为浅层的排名，以及在更严格的裁剪阈值下评估集上的最佳层，都取决于静音。提取 WavLM Base 特征时应避免零填充，使文件的表示仅取决于其自身的采样点：填充会通过卷积编码器的组归一化改变特征，而注意力掩码无法完全消除这种变化。这项建议并非基于性能：在所比较的两种配置中，填充在一些测试集上降低了错误率，在另一些测试集上则提高了错误率。若使用其他自监督学习编码器和训练语料库重复分析，并测量采用更严格裁剪后残留的静音，就能了解这些发现的普适程度。

声明

资助。 本研究未获得公共、商业或非营利部门资助机构的任何专项资助。

利益冲突。 作者声明不存在任何利益冲突。

伦理审批和参与同意。 不适用。本研究仅使用公开可用的 ASVspoof 2019 Logical Access 和 In-the-Wild 数据集；未从人类参与者处收集任何新数据。

出版同意。 不适用。
材料可用性。 不适用。
数据可用性。 ASVspoof 2019 Logical Access 数据库可公开获取，网址为 https://datashare.ed.ac.uk/handle/10283/3336。In-the-Wild 数据集可在以下网址获取：https://deepfake-total.com/in the wild。
代码可用性。 本文图表及所报告数值背后的代码、配置文件和结果表可在 https://github.com/Nooralabd/silence-shortcut-wavlm-layers（版本 v1.0）根据 MIT 许可证获取。

生成式人工智能的使用。 在准备这项工作期间，作者使用了 Claude Opus 5.5（Anthropic），以改进手稿的语言和可读性、组织并记录代码、支持文献检索，以及检查手稿的内部一致性。研究设计、实验、分析和结论均为作者自行完成。在使用该工具后，作者根据需要对内容进行了审阅和编辑，并对出版物的内容承担全部责任。

作者贡献。 N.E.：概念化、方法论、软件、验证、正式分析、调查、数据整理、可视化、撰写初稿。M.E.A.：概念化、方法论、监督、项目管理、撰写审阅与编辑。

附录 A 均值池化下裁剪的影响

表 A1 裁剪的影响（EER 百分点，均值池化），以及 95% 说话人级配对自助法区间。不包含零的区间以粗体显示。评估：ASVspoof 2019 LA 评估集；ITW：In-the-Wild

Set	L	Δ_{test}	Δ_{train}	Δ_{total}	I
评估	1	**+19.88 [17.43, 22.57]**	**−6.83 [−8.75, −4.90]**	**+13.05 [11.30, 14.96]**	**−15.74 [−17.75, −13.86]**
	2	**+13.27 [11.58, 14.93]**	**−3.82 [−5.30, −2.16]**	**+9.45 [8.25, 10.85]**	**−11.74 [−13.13, −10.16]**
	3	**+10.17 [8.81, 11.63]**	**−2.93 [−4.17, −1.88]**	**+7.24 [6.22, 8.39]**	**−8.80 [−10.11, −7.73]**
	4	**+6.60 [5.71, 7.57]**	**−1.73 [−2.63, −0.90]**	**+4.87 [4.03, 5.80]**	**−6.08 [−6.98, −5.21]**
	5	**+7.81 [6.69, 8.95]**	**−3.98 [−4.91, −2.99]**	**+3.84 [3.31, 4.49]**	**−7.70 [−8.71, −6.72]**
	6	**+8.32 [6.82, 9.81]**	**−5.47 [−6.66, −4.29]**	**+2.86 [2.15, 3.51]**	**−7.50 [−8.80, −6.28]**
	7	**+9.16 [7.27, 10.91]**	**−5.47 [−6.87, −4.15]**	**+3.69 [2.79, 4.56]**	**−8.55 [−9.88, −7.16]**
	8	**+12.78 [10.80, 14.72]**	**−7.50 [−9.05, −5.93]**	**+5.28 [4.16, 6.27]**	**−11.60 [−13.31, −9.80]**
	9	**+13.18 [11.33, 14.84]**	**−7.95 [−9.14, −6.61]**	**+5.23 [4.02, 6.29]**	**−12.16 [−13.56, −10.48]**
	10	**+13.03 [11.10, 14.93]**	**−7.65 [−9.06, −6.19]**	**+5.37 [4.22, 6.43]**	**−11.64 [−13.24, −10.05]**
	11	**+15.32 [13.04, 17.49]**	**−9.80 [−11.69, −7.84]**	**+5.52 [4.41, 6.73]**	**−14.69 [−16.64, −12.59]**
	12	**+14.83 [12.46, 17.17]**	**−8.56 [−10.63, −6.39]**	**+6.27 [4.94, 7.63]**	**−14.24 [−16.26, −12.12]**
ITW	1	+0.50 [−1.83, 2.35]	−1.60 [−7.98, 6.51]	−1.10 [−7.43, 6.39]	**−2.12 [−3.70, −0.12]**
	2	−0.53 [−2.20, 1.06]	−3.54 [−8.37, 1.56]	−4.06 [−8.20, 0.64]	**−1.46 [−2.76, −0.29]**
	3	−0.31 [−1.91, 0.94]	−2.90 [−6.21, 0.86]	−3.21 [−6.91, 0.37]	**−3.25 [−5.19, −1.42]**
	4	−0.87 [−2.33, 0.35]	−1.17 [−4.13, 2.10]	−2.04 [−5.07, 0.86]	**−1.32 [−2.04, −0.47]**
	5	−0.88 [−2.43, 0.50]	**−3.53 [−5.90, −0.77]**	**−4.41 [−6.91, −2.18]**	**−2.29 [−3.06, −1.06]**
	6	−1.02 [−2.71, 0.32]	**−4.86 [−6.93, −2.19]**	**−5.89 [−7.78, −3.86]**	−0.95 [−1.92, 0.41]
	7	−0.25 [−1.52, 0.93]	**−3.07 [−5.58, −0.12]**	**−3.32 [−5.64, −0.99]**	**−1.20 [−2.19, −0.18]**
	8	+0.44 [−0.42, 1.45]	**−5.41 [−8.26, −2.59]**	**−4.97 [−7.91, −2.29]**	0.00 [−0.97, 0.72]
	9	+0.26 [−0.72, 1.27]	**−4.55 [−6.36, −2.71]**	**−4.28 [−6.28, −2.18]**	**−1.09 [−1.91, −0.05]**
	10	+0.48 [−1.09, 1.88]	−1.65 [−4.32, 1.83]	−1.17 [−3.88, 1.60]	−0.75 [−1.76, 0.50]
	11	−0.08 [−1.86, 1.42]	+0.29 [−2.57, 3.41]	+0.20 [−2.34, 2.63]	+0.09 [−1.09, 1.49]
	12	−0.30 [−2.04, 1.34]	+0.19 [−2.89, 4.28]	−0.12 [−2.62, 3.24]	+1.03 [−0.49, 2.62]

表 A2 设计中剩余的两种影响 2×2 （EER 百分点），附有 95% 的说话人级配对自助法区间，使用与表 5 相同的重采样：在测试音频干净时，使用裁剪音频进行训练， $\Delta^{\mathrm{C}}_{\mathrm{train}}$，以及裁剪使用裁剪音频训练的模型的测试音频， $\Delta^{\mathrm{T}}_{\mathrm{test}}$ （公式（7）–（8））。不包含零的区间以粗体显示

		$\Delta^{\mathrm{C}}_{\text{训练集}}$		$\Delta^{\mathrm{T}}_{\mathrm{test}}$	
Set	L	统计数据	mean	统计数据	mean
评估	1	**+8.90 [7.43, 10.43]**	**+8.91 [7.56, 10.56]**	**+3.92 [2.80, 5.00]**	**+4.13 [3.09, 5.34]**
	2	**+7.51 [6.36, 8.87]**	**+7.92 [6.69, 9.31]**	**+1.60 [0.83, 2.40]**	**+1.53 [0.78, 2.42]**
	3	**+5.49 [4.55, 6.54]**	**+5.87 [4.93, 7.02]**	**+1.56 [0.97, 2.14]**	**+1.37 [0.73, 1.98]**
	4	**+3.89 [3.14, 4.70]**	**+4.35 [3.52, 5.18]**	**+0.62 [0.18, 1.08]**	**+0.52 [0.03, 1.03]**
	5	**+3.49 [2.85, 4.15]**	**+3.72 [3.03, 4.48]**	+0.17 [−0.36, 0.61]	+0.11 [−0.45, 0.72]
	6	**+2.10 [1.59, 2.57]**	**+2.03 [1.50, 2.49]**	**+0.79 [0.37, 1.30]**	**+0.82 [0.41, 1.26]**
	7	**+3.10 [2.57, 3.75]**	**+3.08 [2.51, 3.68]**	+0.39 [−0.42, 1.24]	+0.61 [−0.10, 1.29]
	8	**+4.47 [3.79, 5.08]**	**+4.10 [3.40, 4.76]**	+0.42 [−0.57, 1.41]	**+1.18 [0.11, 2.17]**
	9	**+4.35 [3.56, 5.20]**	**+4.21 [3.36, 4.94]**	+0.62 [−0.36, 1.55]	+1.02 [−0.01, 2.07]
	10	**+4.37 [3.58, 5.11]**	**+3.99 [3.23, 4.69]**	+0.80 [−0.21, 1.78]	**+1.39 [0.31, 2.45]**
	11	**+4.92 [3.83, 6.01]**	**+4.89 [3.96, 5.85]**	+0.33 [−0.45, 1.18]	+0.63 [−0.15, 1.40]
	12	**+5.85 [4.77, 6.94]**	**+5.69 [4.61, 6.69]**	+0.19 [−0.73, 1.28]	+0.59 [−0.31, 1.70]
ITW	1	−0.26 [−6.36, 6.59]	+0.52 [−5.90, 7.89]	**−1.39 [−2.35, −0.56]**	**−1.62 [−2.71, −0.75]**
	2	−2.54 [−6.80, 2.34]	−2.08 [−6.55, 3.35]	**−1.93 [−2.97, −0.79]**	**−1.99 [−3.37, −0.94]**
	3	−0.88 [−5.42, 4.14]	+0.35 [−3.90, 5.24]	**−3.38 [−5.53, −1.56]**	**−3.56 [−5.88, −1.63]**
	4	−1.32 [−4.70, 2.09]	+0.14 [−2.69, 3.26]	**−1.85 [−2.79, −0.98]**	**−2.19 [−3.23, −1.25]**
	5	−1.41 [−3.76, 1.43]	−1.24 [−3.89, 1.30]	**−2.53 [−3.80, −1.46]**	**−3.17 [−4.47, −1.67]**
	6	**−3.93 [−6.03, −1.46]**	**−3.92 [−6.03, −1.64]**	**−1.60 [−2.73, −0.64]**	**−1.97 [−3.22, −0.73]**
	7	−1.62 [−4.34, 1.04]	−1.87 [−4.55, 1.01]	−0.71 [−1.46, 0.07]	**−1.45 [−2.37, −0.53]**
	8	**−4.16 [−6.86, −1.36]**	**−5.41 [−8.26, −2.69]**	**+0.95 [0.52, 1.45]**	+0.44 [−0.07, 0.80]
	9	**−2.53 [−4.83, −0.37]**	**−3.45 [−5.66, −1.14]**	−0.08 [−0.48, 0.48]	**−0.83 [−1.68, −0.08]**
	10	−0.67 [−3.22, 1.99]	−0.90 [−3.61, 2.11]	+0.15 [−0.31, 0.69]	−0.27 [−0.96, 0.40]
	11	−0.43 [−3.01, 2.35]	+0.19 [−2.36, 2.83]	+0.28 [−0.36, 0.87]	+0.01 [−0.63, 0.55]
	12	−0.17 [−3.00, 2.97]	−0.85 [−3.43, 2.50]	**+0.91 [0.12, 1.73]**	**+0.73 [0.01, 1.51]**

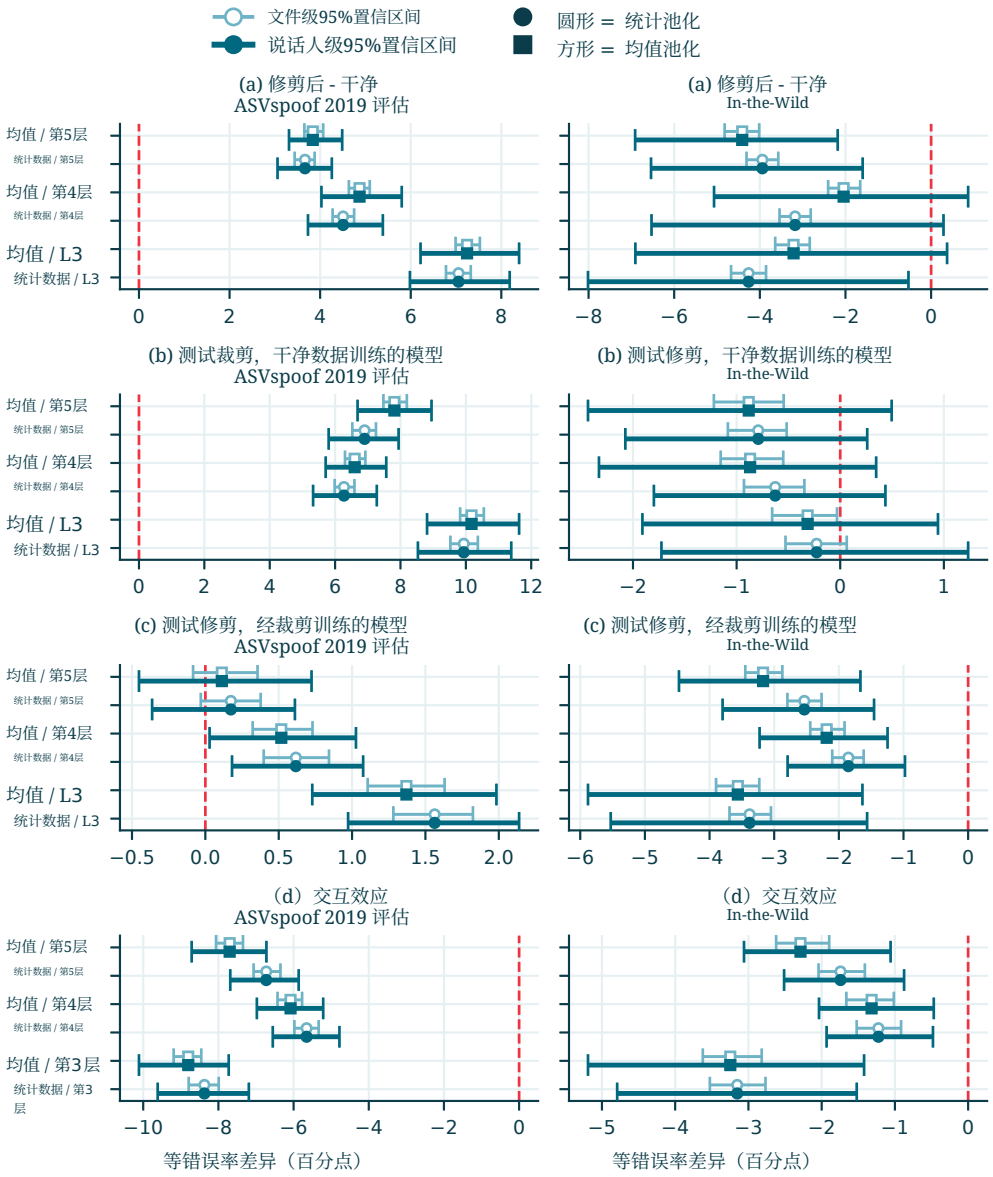

图 A1 六种配置（第3–5层、均值池化和统计池化）在评估集（左）和 In-the-Wild（右）上裁剪的影响：(a) 修剪后减去干净（Δ_{total}）；(b) 对在干净音频上训练的模型的测试音频进行裁剪（Δ_{test}）；(c) 对在裁剪音频上训练的模型的测试音频进行裁剪（$\Delta^{\text{T}}_{\text{test}}$）；以及 (d) 交互效应 I。圆形：统计池化；方形：均值池化；红色虚线表示零值。所有区间均为配对的、按类别分层的百分位数自助法置信区间（1,000次重采样），针对种子平均等错误率（每种子等错误率，五个随机种子的平均值）；浅色：文件层级，深色：说话人级

附录 B 秩一致性的区间构造

表B1 同一组 1,000 次重采样中，干净层级排序与修剪后层级排序之间的 Kendall's τ，以及说话人层面的 95 % 百分位数区间和偏差校正加速区间； z_0：偏差校正常数（并列计为二分之一）； a：加速量（类别内逐一排除说话人的刀切法）；并列：与观测到的 τ 相等的重采样所占比例； p：针对 $n = 12$ 的精确双侧 p 值

阈值	Set	池化	τ	百分位数	BCa	z_0	a	Ties	p
40 分贝	Eval	统计数据	0.58	[0.39, 0.64]	[0.52, 0.70]	0.86	0.019	18%	0.009
	Eval	mean	0.58	[0.42, 0.64]	[0.52, 0.67]	0.89	−0.007	15%	0.009
	ITW	统计数据	0.73	[0.42, 0.85]	[0.58, 0.93]	0.49	0.017	10%	0.0005
	ITW	mean	0.76	[0.39, 0.82]	[0.67, 0.91]	0.99	0.027	8%	0.0002
20 分贝	Eval	统计数据	0.48	[0.27, 0.52]	[0.45, 0.55]	1.21	0.010	10%	0.031
	Eval	mean	0.39	[0.27, 0.48]	[0.27, 0.48]	−0.09	−0.015	23%	0.086
	ITW	统计数据	0.52	[0.27, 0.82]	[0.01, 0.70]	−0.46	−0.025	10%	0.021
	ITW	mean	0.58	[0.27, 0.79]	[0.36, 0.83]	0.24	−0.003	11%	0.009
15 分贝	Eval	统计数据	0.42	[0.27, 0.52]	[0.30, 0.55]	0.22	−0.012	19%	0.063
	Eval	mean	0.36	[0.27, 0.48]	[0.21, 0.45]	−0.39	−0.016	21%	0.116
	ITW	统计数据	0.64	[0.27, 0.79]	[0.48, 0.93]	0.71	0.024	6%	0.003
	ITW	mean	0.48	[0.24, 0.76]	[0.18, 0.70]	−0.23	−0.016	9%	0.031

参考文献

[1] Tan, X., Qin, T., Soong, F., Liu, T.-Y.：神经语音合成综述。arXiv 预印本 arXiv:2106.15561（2021）。https://doi.org/10.48550/arXiv.2106.15561

[2] Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C.Y., Zhao, Y.：音频深度伪造检测：综述。arXiv 预印本 arXiv:2308.14970（2023）。https://doi.org/10.48550/arXiv.2308.14970

[3] Kinnunen, T., Delgado, H., Evans, N., Lee, K.A., Vestman, V., Nautsch, A., Todisco, M., Wang, X., Sahidullah, M., Yamagishi, J., Reynolds, D.A.：语音欺骗对策与自动说话人验证的串联评估：基础。IEEE/ACM 音频、语音与语言处理汇刊 **28**，2195–2210（2020）https://doi.org/10.1109/TASLP.2020.3009494

[4]Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N.,Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K.A., 等：ASVspoof 2019：合成、转换和重放语音的大规模公共数据库。计算机语音与语言**64**，101114（2020）https://doi.org/10.1016/j.csl.2020.101114

[5] Baevski, A., Zhou, H., Mohamed, A., Auli, M.：wav2vec 2.0：一种用于语音表示自监督学习的框架。载于：神经信息处理系统进展，第33卷，第12449–12460页（2020）

[6] Hsu,W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.：HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习。IEEE/ACM音频、语音与语言处理汇刊**29**，3451–3460（2021）https://doi.org/10.1109/TASLP.2021.3122291

[7] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F.：WavLM：面向全栈语音处理的大规模自监督预训练。IEEEJournalofSelectedTopicsinSignalProcessing**16**(6)，1505–1518（2022）https://doi.org/10.1109/JSTSP.2022.3188113

[8] Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., Auli, M.：XLS-R：面向大规模的自监督跨语言语音表征学习。载于：Interspeech 2022论文集，第2278–2282页（2022）。https://doi.org/10.21437/Interspeech.2022-143[9]Wang, X., Yamagishi, J.：研究用于语音欺骗对策的自监督前端。载于：Odyssey 2022：说话人和语言识别研讨会论文集，第100–106页（2022）。https://doi.org/10.21437/Odyssey.2022-14

[10] Guo, Y., Huang, H., Chen, X., Zhao, H., Wang, Y.：结合自监督WavLM和多融合注意力分类器进行音频深度伪造检测。载于：IEEE国际声学、语音与信号处理会议（ICASSP）论文集，第12702–12706页（2024）。https://doi.org/10.1109/ICASSP48485.2024.10447923[11] El Kheir, Y., Samih, Y., Maharjan, S., Polzehl, T., M¨oller, S.：针对音频深度伪造检测的自监督学习模型全面逐层分析。载于：计算语言学协会论文集：NAACL 2025 Findings，第4070–4082页（2025）。https://doi.org/10.18653/v1/2025.findings-naacl.227[12] Pan, Z., Liu, T., Sailor, H.B., Wang, Q.：对预训练语音模型的隐藏嵌入进行注意力合并以实现反欺骗检测。载于：Interspeech 2024论文集，第2090–2094页（2024）。https://doi.org/10.21437/Interspeech.2024-1472[13] Sinha, S., Dey, S., Saha, G.：通过层条件嵌入融合改进用于音频欺骗检测的自监督学习模型。《计算机语音与语言》86，101599（2024）https://doi.org/10.1016/j.csl.2023.101599[14] Beheshti, M., Rostami, M., Chen, B.：从自监督语音模型中进行探测引导层选择，以实现可泛化的音频深度伪造检测。arXiv预印本 arXiv:2606.30791（2026）。https://doi.org/10.48550/arXiv.2606.30791[15] Serrano, P., Duroselle, R., Angulo, F., Bonastre, J.-F., Boeffard, O.：通过层选择和基于SSL的对策融合改进域外音频深度伪造检测。arXiv预印本 arXiv:2509.12003（2025）。https: // doi.org/10.48550/arXiv.2509.12003[16] M¨uller, N.M., Dieckmann, F., Czempin, P., Canals, R., B¨ottinger, K., Williams, J.：语音如银，静音如金：ASVspoof训练的模型究竟学到了什么？载于：2021年自动说话人验证与欺骗对策挑战赛论文集，第55–60页（2021）。https://doi.org/10.21437/ ASVspoof.2021-9 [17] Zhang, Y., Li, Z., Lu, J., Hua, H., Wang, W., Zhang, P.：静音对语音反欺骗的影响。《IEEE/ACM音频、语音与语言处理汇刊》31，3374–3389（2023）https://doi.org/10.1109/TASLP.2023.3306711[18] Liu, X.,Wang, X.,Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A., Lee, K.A.：ASVspoof 2021：迈向真实环境中的伪造语音与深度伪造语音检测。《IEEE/ACM音频、语音与语言处理汇刊》31，2507–2522（2023）https://doi.org/ 10.1109/TASLP.2023.3285283[19]Shim, H.-j., Gonzalez Hautam¨aki, R., Sahidullah, M., Kinnunen, T.：如何构建完美和比抛硬币猜测更差的欺骗对策：关于捷径学习的警示。载于：Interspeech 2023论文集，第785–789页（2023）。https://doi.org/10.21437/Interspeech.2023-1901

[20] Wang, L., Yu, L., Zhang, Y., Xie, H.: 结合数据增强与多任务元学习，实现针对静音裁剪的泛化语音欺骗检测。IEEE/ACM音频、语音与语言处理汇刊 **32**, 3296–3310(2024) https://doi.org/10.1109/TASLP.2024.3414340[21] Xiao, Y., Vu, N.T.: 利用XLS- R进行伪造音频检测的逐层决策融合。载于：Interspeech 2025论文集，第5618–5622页（2025）。https://doi.org/10.21437/Interspeech.2025-1543[22] M¨uller,N.M., Czempin, P., Dieckmann, F., Froghyar, A., B¨ottinger, K.: 音频深度伪造检测能否实现泛化？载于：Interspeech 2022论文集，第2783–2787页（2022）。https://doi.org/10.21437/Interspeech.2022-108[23] Shim, H.-j., Sahidullah, M., Jung, J.-w., Watanabe, S., Kinnunen, T.: 超越静音：通过损失和非对称方法分析音频反欺骗中的偏差。载于：SynData4GenAI 2024论文集，第56–60页（2024）。https://doi.org/10.21437/SynData4GenAI.2024-12[24] Wu,Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilci, C., Sahidullah, M., Sizov, A.: ASVspoof 2015：首届自动说话人验证欺骗与对策挑战赛。载于：Interspeech 2015论文集，第2037–2041页（2015）。https://doi.org/10.21437/Interspeech.2015-462[25] Lavrentyeva, G., Novoselov, S., Tseren,A., Volkova, M., Gorlanov, A., Kozlov, A.: 面向ASVspoof2019挑战赛的STC反欺骗系统。载于：Interspeech 2019论文集，第1033–1037页（2019）。https://doi.org/10.21437/Interspeech.2019-1768[26] Jung, J.-w., Heo, H.-S., Tak, H., Shim, H.-j., Chung, J.S., Lee, B.-J., Yu, H.- J., Evans, N.: AASIST：利用集成频谱时序图注意力网络进行音频反欺骗。载于：IEEE国际声学、语音与信号处理会议（ICASSP）论文集，第6367–6371页（2022）。https://doi.org/10.1109/ICASSP43922.2022.9747766[27]Zhang, Y., Lu, J., Shang, Z., Wang, W., Zhang, P.: 利用AASIST2提升短话语反欺骗性能。载于：IEEE国际声学、语音与信号处理会议（ICASSP）论文集，第11636–11640页（2024）。https://doi.org/10.1109/ICASSP48485.2024.10448049[28] Zhang, Q., Wen, S., Hu, T.: 利用自监督XLS-R和SLS分类器进行音频深度伪造检测。载于：第32届ACM多媒体国际会议论文集，第6765–6773页（2024）。https://doi.org/10.1145/3664647.3681345[29]Rosello, E., Gomez-Alanis, A., Gomez, A.M., Peinado, A.M.: 一种基于Conformer的分类器，用于反欺骗中的变长语音处理。载于：Interspeech 2023论文集，第5281–5285页（2023）。https://doi.org/10.21437/Interspeech.2023-1820

[30] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X向量：用于说话人识别的鲁棒深度神经网络嵌入。载于：IEEE国际声学、语音与信号处理会议（ICASSP）论文集，第5329–5333页（2018）。https://doi.org/10.1109/ICASSP.2018.8461375[31]Okabe, K., Koshinaka, T., Shinoda, K.: 用于深度说话人嵌入的注意力统计池化。载于：Interspeech 2018论文集，第2252–2256页（2018）。https://doi.org/10.21437/Interspeech.2018-993[32] Stourbe, T., Miara, V., Lepage, T., Dehak, R.: 探索用于语音欺骗和深度伪造检测的WavLM后端。载于：ASVspoof 2024研讨会论文集，第72–78页（2024）。https://doi.org/10.21437/ASVspoof.2024-11[33] Pasad,A., Chou,J.-C.,Livescu, K.:自监督语音表示模型的逐层分析。载于：2021年IEEE自动语音识别与理解研讨会（ASRU）论文集，第914–921页（2021）。https://doi.org/10.1109/ASRU51503.2021.9688093[34] Ali, H., Adupa, N.S., Subramani, S., Malik, H.: 面向音频深度伪造检测的自监督语音模型SUPERB风格基准。载于：IEEE国际声学、语音与信号处理会议（ICASSP）论文集，第16512–16516页（2026）。https://doi.org/10.1109/ICASSP55912.2026.11460421[35] Zhu, Y., Powar, S., Falk, T.H.: 表征通用语音表示的时间动态，以实现泛化深度伪造检测。载于：IEEE国际声学、语音与信号处理研讨会（ICASSPW）论文集，第139–143页（2024）。https://doi.org/10.1109/ICASSPW62465.2024.10626010[36] Mallat, M., Panariello, M., Todisco, M., Evans, N., Larcher, A.: 解读用于欺骗检测的自监督学习表征：WavLM研究。载于：2026年第14届国际生物识别与取证研讨会（IWBF）论文集，第1–6页（2026）。https://doi.org/10.1109/IWBF68042.2026.11558152[37] Rubio, S., Bello, P., Ribas, D., Miguel, A., Lleida, E., Ortega, A.: 一种基于干预的框架，用于诊断欺骗对策中的捷径。载于：Odyssey 2026论文集，第1–8页（2026）。https://doi.org/10.21437/Odyssey.2026-1[38] Wang, H., Fan, C., Zhou, J., Yi, J., Kwan, H.K.: 一种多子空间注意力方法，用于静音裁剪条件下的鲁棒语音欺骗检测。载于：《人机语音通信》。计算机与信息科学通信，第2662卷，第328–339页。Springer，新加坡（2026）。https://doi.org/10.1007/978-981-95-5382-2 26[39] Xiao, Y., Lin, C.-W., Wang, X., El Kheir, Y., Das, A., Polzehl, T., M¨oller, S., Vu, N.T.: 证据子空间投影：衡量证据对自监督语音模型深度伪造检测的解释程度。arXiv预印本

arXiv:2607.11538（2026）。https://doi.org/10.48550/arXiv.2607.11538

[40] Kinnunen, T., Lee, K.A., Delgado, H., Evans, N., Todisco, M., Sahidullah, M., Yamagishi, J., Reynolds, D.A.：t-DCF：用于评估语音欺骗对策与自动说话人验证串联性能的检测成本函数。载于：Odyssey 2018：说话人和语言识别研讨会论文集，第312–319页（2018）。https://doi.org/10.21437/Odyssey.2018-44[41] Efron, B.：更优的自助法置信区间。《美国统计学会杂志》**82**（397），171–185（1987）https://doi.org/10.1080/01621459.1987.10478410[42] Pascu, O., Stan, A., Oneata, D., Oneata, E., Cucu, H.：利用自监督表征实现泛化能力更强且经过校准的音频深度伪造检测。载于：Interspeech 2024论文集，第4828–4832页（2024）。https://doi.org/10.21437/Interspeech.2024-1302