

离散最优传输是一种强大的音频对抗攻击

A. Selitskiy,¹ A. Shahriyar,² and J. Prakashan²

¹University of Rochester, ²Rochester Institute of Technology

ABSTRACT

在本文中，我们展示了离散最优传输（DOT）是一种有效的黑箱对抗性攻击，针对现代音频防欺骗对策（CM）。我们的攻击操作作为后处理、分布对齐步骤：生成语音的帧级 WavLM 嵌入通过熵 OT 和 top- k 巴氏投影与一个不配对的真实池对齐，然后使用神经声码器进行解码。在 AASIST 基准测试上的 ASVspoof2019 和 ASVspoof5 数据集上评估 DOT，在各个数据集中持续产生高等错误率（EER），并在 CM 微调后仍保持竞争力，并且优于几种传统的跨数据集传输攻击。消融分析突显了声码器重叠的实际影响。结果表明，分布级对齐是对部署的 CM 的一种强大且稳定的攻击面。

Index Terms— 最优传输，对抗攻击，ASVspoof

1. 介绍

基于向量嵌入的语音转换（VC）使用 WavLM 模型 [1] 从嵌入空间中的一个简单的 k -最近邻（ k NN）映射开始 [2]。后续工作用最优化传输（OT）替换了 k NN，改进了源分布与真实目标分布之间的对齐，并提升了转换质量 [3, 4]。特别是最近在生成的音频上应用了带有重心投影的离散 OT（DOT）作为后处理域对齐步骤，即使是在强反制措施（CMs）下，也得到了接近真实录音的成绩 [4]。类似地，在反欺骗中的分布对比方面，也在防御方面探索了 OT 的应用 [5]。

大量研究表明，自动语音识别（ASR）和说话人验证（ASV）管道易受对抗性示例的影响，包括基于优化的攻击如 [6]（另见 [7, section 7]）。同时，ASVspoof 挑战赛建立了标准化的数据集和协议（例如，ASVspoof2019 [8]；ASVspoof5 [7]），用于

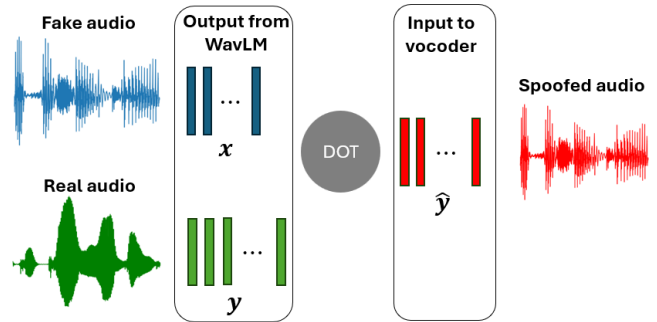


Fig. 1. 离散 OT 对抗攻击系统概述。

训练和评估欺骗 CMs。在现代 CMs 中，AASIST 系统 [9] 及其修改版本利用频谱-时间图注意力来捕捉多种欺骗伪影，并已成为深度伪造/反欺骗检测的强基线。尽管取得了这些进展，跨数据集和生成模型的可转移性 仍然是一个开放挑战。

我们研究基于离散 OT 的语音转换是否是针对最先进的 CM 的一种强大黑箱攻击（图 1）。我们的方法通过熵 DOT 和余弦成本将 WavLM 帧嵌入对齐到一个未配对的真实样本池中，应用前 k 巴氏中心映射，并使用神经声码器进行重建。与梯度攻击不同，DOT 诱导了一个朝向真实区域的分布移位，使得在无需 CM 内部信息的情况下实现跨数据集迁移。**我们的贡献：**

- **DOT 作为对抗攻击。**我们将离散 OT + 重心投影 VC 形式化为一个黑盒，分布对齐攻击。
- **传输和微调鲁棒性。**DOT 在 ASVspoof2019 和 ASVspoof5 数据集上保持强大，并且在 CM 微调后仍具有竞争力，超越了几种传统攻击。
- **消融实验和实际因素。**我们证明了带有 CM 训练数据的声码器重叠调节了攻击强度，这对攻击设

计和 CM 训练都有影响。

2. 问题设定和威胁模型

我们考虑一个音频助手/ASV 流水线, 其欺骗 CM 是在 ASVspoof 协议下训练的 [8, 7]。给定波形 x , CM 输出一个分数 $f(x)$; 一个阈值 τ 产生一个二元 (真/欺骗) 决策。

对手目标。 将生成/转换的输入 x_{gen} 转变为一个攻击样本 y , 使其 (i) 被 CM 识别为被接受为真实无误的并且 (ii) 保持可理解性/自然度以避免人类怀疑。

对手的知识 and 访问权限。 我们假设一个黑箱 CM: 没有梯度或内部信息; 访问分数/标签是可选的, 不需要。对手可以使用任何上游 TTS/VC 系统合成 x_{gen} , 并且有一个未配对来自目标领域或类似代理的真实语音池, 这反映了现实部署和标准对抗音频实践。

对抗能力 (点破坏攻击)。 对手将 x_{gen} 嵌入帧级向量 $X = \{x_i\}_{i=1}^M$, 嵌入一个未配对的真实池 $Y = \{y_j\}_{j=1}^N$, 在余弦代价 $c(x, y) = 1 - \cos(x, y)$ 下计算一个耦合矩阵熵离散 $\text{OT}\gamma \in \mathbf{R}^{M \times N}$, 应用顶点- k 重心投影获得传输嵌入 $\hat{Y} = \{\hat{y}_i\}_{i=1}^M$, 并通过神经声码器重构波形 $\tilde{x}$ 。该管道仅使用通用的预训练组件和不对的语音。

感知约束。 该攻击旨在与 x_{gen} 保持最小的感知差异, 同时实现 CM 规避。

我们遵循 ASVspoof 协议 [8, 7]。成功通过在 $\tilde{x}$ 上的更高的 EER (或更高的误接受率) 来证明, 包括在跨数据集迁移下和 CM 微调之后的表现优于基线。我们还分析了调节攻击强度的实际因素——声码器重叠、top- k 和目标持续时间 (参见第 5 节)。

3. 方法论: 离散 OT 攻击

3.1. 离散最优传输和重心投影

设 $(X, \mathcal{P}, \mathbf{P})$ 和 $(Y, \mathcal{Q}, \mathbf{Q})$ 是两个概率空间。记 $\Pi(\mathbf{P}, \mathbf{Q})$ 为在积空间 $(X \times Y, \mathcal{P} \otimes \mathcal{Q}, \pi)$ 上的所有联合分布, 其边缘分布分别为 $\mathbf{P}$ 和 $\mathbf{Q}$ 。在离散情形中假设 X 中有 M 个向量, 在 Y 中有 N 个向量, 其概率质量分别为 $p_i = \mathbf{P}(x_i)$ 和 $q_j = \mathbf{Q}(y_j)$ 。联合分布

$\pi(x, y)$ 被表示为一个非负矩阵 γ , 与 $\gamma_{ij} = \pi(x_i, y_j)$, $i = 1, \dots, M$ 和 $j = 1, \dots, N$. 相关。

最优传输 (OT) 的目标是找到已知为康托罗维奇计划, 的联合分布 $\pi \in \Pi(\mathbf{P}, \mathbf{Q})$, 该分布最小化期望传输成本。

$$\sum_i \sum_j \gamma_{ij} c(x_i, y_j) \rightarrow \inf_{\gamma_{ij}}, \quad (1)$$

受边际约束条件的限制:

$$p_i = \sum_{j=1}^N \gamma_{ij} \quad \text{and} \quad q_j = \sum_{i=1}^M \gamma_{ij}. \quad (2)$$

给定一个解 γ , 可以通过重心投影: 定义一个传输映射

$$T(x_i) = \sum_{j=1}^N \tilde{\gamma}_{ij} y_j, \quad \text{where } \tilde{\gamma}_{ij} = \frac{\gamma_{ij}}{p_i}. \quad (3)$$

此变换可以解释为条件期望 $\mathbf{E}[y|x = x_i]$ 。

3.2. 对抗攻击

我们遵循离散 OT 语音转换框架的 [4]。令一个音频记录由一系列嵌入表示 $\mathbf{x} = [x_1, x_2, \dots, x_M]$ 。这里, x_i , $i = 1, \dots, M$, 是使用 WavLM Large 预训练模型 [1] 获得的嵌入。该模型将每 25 毫秒的音频编码为一个 1024 维的向量嵌入, 步长为 20 毫秒。

对于每一对分别来自源说话人和目标说话人的音频录音 $(\mathbf{x}, \mathbf{y})$, 我们提取它们的向量表示: $\mathbf{x}$ 和 $\mathbf{y}$, 使用 $x_i, y_j \in \mathbf{R}^{1024}$ 。

由于说话人嵌入 $X = \{x_i\}_{i=1}^M$ 和 $Y = \{y_j\}_{j=1}^N$ 的底层分布未知, 我们使用经验分布: $\mathbf{P}(x_i) = \frac{1}{M}$ 和 $\mathbf{P}(y_j) = \frac{1}{N}$ 。注意到目标说话人的录音较长会更好; 因此, Y 可以包含来自多个音频样本的嵌入。

目标是通过将 OT 形式嵌入 $\mathbf{x}$ 应用于真实样本 $\mathbf{y}$ 来增强生成的音频 x_{gen} 。换句话说, 我们希望定义一个映射 (传输映射) T , 使得 $T(\mathbf{x}) \approx \mathbf{y}$, 其中 $\mathbf{y}$ 具有与 $\mathbf{Q}$ 相似的分佈 (但它不一定是 Y 的元素)。在离散 OT 方法中, 我们计算给定边缘分布和成本函数的耦合矩阵 γ 。对于每个 x_i , 我们将目标嵌入 y_j 按降序排列, 以 γ_{ij} 表示, 排序后的向量记为 $y_j^{\text{ot}(i)}$ 。每行 (固定 i) 的排序耦合权重记为 $\gamma_{ij}^{\text{sort}}$ 。

我们定义映射 $\hat{T}$ 为 OT 映射在顶部 k 向量上的重心投影,

$$x_i \mapsto \hat{y}_i = \sum_{j=1}^k \tilde{\gamma}_{ij}^{sort} y_j^{ot(i)}, \quad \tilde{\gamma}_{ij}^{sort} = \frac{\gamma_{ij}^{sort}}{\sum_{s=1}^k \gamma_{is}^{sort}}. \quad (4)$$

请注意, 该公式仅在 $k = N$ 时与 (3) 相符。我们使用了 $k = 5$, 正如在 [4] 中所示, 转换的质量对于 $3 \leq k \leq 10$, 变化不大, 但对于较大的 k , 则会下降。经过变换 $\mathbf{x} \mapsto \hat{\mathbf{y}}$, 后, 我们使用 HiFi-GAN 语音码器 (参见第 4 节) 将预测的嵌入 $\hat{\mathbf{y}}$ 转换回波形 $\hat{\mathbf{y}}$ 。

这种攻击背后的直觉很简单: 近似映射 $\hat{T}$ 将生成语音的经验分布向目标 (真实) 分布移动。由于防御系统被训练成拒绝合成的分布并接受真实的语音, 将生成的分布更接近真实分布可以降低检测性能——即产生一种强大且与数据集无关的对抗效果。

3.3. 攻击变种

作为集合 X , 我们使用了来自 ASVspoof2019 数据集 (验证部分) 的生成音频。作为目标空间 Y , 我们研究了两个选项: 半消声室录音和多样化的录音条件。结果, 我们考虑了以下攻击。

OT₁: 作为目标空间 Y , 我们使用了基于 VCTK [10] 构建的 ASVspoof2019 数据集的真实录音。

OT₂: 作为目标空间 Y , 我们使用了 LibriSpeech train-clean-100 数据集 [11], 因为 ASVspoof5 使用了包含 Libri Speech 的 LibriVox 数据。我们选择前 40 位说话者 (按说话者 ID 排序), 并为每位说话者提取 10 条随机话语, 并按持续时间对它们进行排序。

先前的研究 (参见例如 [4]) 注意到, 语音转换的质量依赖于目标数据的长度。使用 LibriSpeech 数据集可以将同一说话人的几个话语连接起来。评估显示, 使用 VCTK 或 LibriSpeech 之间的差异不大 (见表 1), 我们决定使用 OT₂ 作为进一步分析的主要攻击方法。

4. 实验设置

数据集。 我们使用三个公共语料库。为了在我们的 DOT 攻击 (OT₂) 中构建真实的靶标池, 我们从

Kaggle [12] 上的 LibriSpeech 清晰版中提取。对于对策 (CM) 评估和跨数据集迁移, 我们使用了可在 [13] 上访问的 ASVspoof2019 及 ASVspoof5 [7]。除非另有说明, 否则我们将使用每个基准的官方训练/开发/评估划分。

嵌入。 我们使用 WavLM 大型版本提取帧级嵌入, 通过 KNN-VC 框架从第六个变换器层获取嵌入。[3]

最优传输。 离散 OT 通过来自潜力库的 Sinkhorn 算法使用熵正则化求解 [14]。使用余弦成本 $c(x, y) = 1 - \cos(x, y)$; 正则化参数设置为 $\epsilon = 0.1$ 。Top- k 重心投影产生传输嵌入具有 $k = 5$ 。

声码器。 我们使用 HiFi-GAN 从传输的嵌入中重建波形, 使用了 KNN-VC 框架提供的实现 [3]。

对策。 对于评估, 我们采用官方的辅助系统智能技术实现 [9] 及其预训练变体。

评估指标 1: 等错误率 我们报告了等错误率 (EER), 即假接受率等于假拒绝率的工作点 (参见, 例如 [15, section 13])。

评估指标 2: 分布相似性。 为了独立于特定的 CM 来量化 DOT 的分布对齐效果, 我们计算了弗雷歇音频距离 (FAD) [16]。我们使用了 torchvggish (v0.2) 用于 VGGish 嵌入 [17], 详情见 [4]。

5. 分析与评估

为了评估, 我们使用了在 ASVspoof2019 数据集上预训练的 AASIST 模型 [9] (表示为 AASIST₂₀₁₉), 以及在 ASVspoof5 上的预训练模型 (表示为 AASIST₅)。

表 1 报告了来自 ASVspoof2019 的生成算法 A18 和 ASVspoof5 中的攻击 A18 (我们用 A18₅ 表示) 以及前一节介绍的两种攻击的等错误率 (EER)。选择算法 A18 和 A18₅ 是因为它们在其各自的数据集中表现出最高的 EER。

Table 1. 最强攻击和提议攻击的 EER↓。

Attack	AASIST ₂₀₁₉	AASIST ₂₀₁₉ ^{FT}	AASIST ₅	AASIST ₅ ^{FT}
A18	2.614	3.141	77.735	44.951
A18 ₅	0.435	1.443	57.933	2.730
OT ₁	11.111	-	7.268	-
OT ₂	7.925	0.216	11.180	12.586

列 $AASIST_{2019}$ 显示了在使用 $AASIST_{2019}$ 评估的 $ASVspoof2019$ 验证子集上的 EER (即, 使用来自 $ASVspoof2019$ 验证集的真实数据)。

列 $AASIST_5$ 提供了使用 $AASIST_5$ 计算的 EER。一个值得注意的观察结果是: 在 $ASVspoof5$ 中提出的最强攻击被预先训练在 $ASVspoof2019$ 上的模型稳健地检测到了。相反, 大多数来自 $ASVspoof2019$ 的方法比 $ASVspoof5$ 中的新攻击具有更高的 EER (也请参见表 2 中的第 (2019) 列和第 (5) 列)。这种不对称性反映了模型/数据集对抗迁移性之间的 [18, 19, 20], 这在以前的语音系统中也观察到了 [21, 22]。

列 $AASIST_{2019}^{FT}$ 显示了在训练集 $ASVspoof2019$ 中包含 OT_2 示例后对 $AASIST_{2019}$ 进行微调的结果。

最后一列报告了使用 OT_2 和 $A18_5$ 数据对 $AASIST_5$ 进行微调后的结果 (详情见第 5.1 节)。

表 2 扩展了这一分析, 包括攻击 A07–A19 ($ASVspoof2019$) 和 A17₅–A31₅ ($ASVspoof5$)。列 (2019) 和 (5) 表示使用 $AASIST_{2019}$ 和 $AASIST_5$ 的基线评估, 而列 (2019_{OT}) 和 (5_{OT}) 显示在对 A07–A19 数据应用最优传输后的结果。

5.1. 微调

由于所有生成方法和来自 $ASVspoof5$ 的攻击在被 $AASIST_{2019}$ 检测时都导致了非常低的等错误率 (EER), 我们仅使用 OT_2 数据对 $AASIST_{2019}$ 进行微调。这些示例是通过对 $ASVspoof2019$ 训练集中生成的音频应用最优传输而获得的。

比较列 (2019_{OT}) 和 (2019_{OT}^{FT}) 在表 2 中 (等同于表 1 中的 $AASIST_{2019}$ 列和 $AASIST_{2019}^{FT}$ 列), 我们可以看到, 经过微调后很容易检测到 OT_2 攻击。A18₅ 的 EER 基本保持不变。

对于遭受强烈 A18 攻击 (也受到 A19 的影响, 但这里我们专注于对抗性攻击) 的 $AASIST_5$, 我们使用了来自 $ASVspoof5$ 评估集的一部分 A18₅ 数据进行微调 (27,000 条录音中的 12,000 条, 其余保留用于评估), 并包含了用于微调 $AASIST_{2019}$ 的 OT_2 训练数据。

$AASIST_5$ 和 $AASIST_5^{FT}$ 的比较 (表 1) 显示, A18₅ 在微调后变得可以很好地检测到。然而, OT_2 攻击仍然保持较高的等错误率。

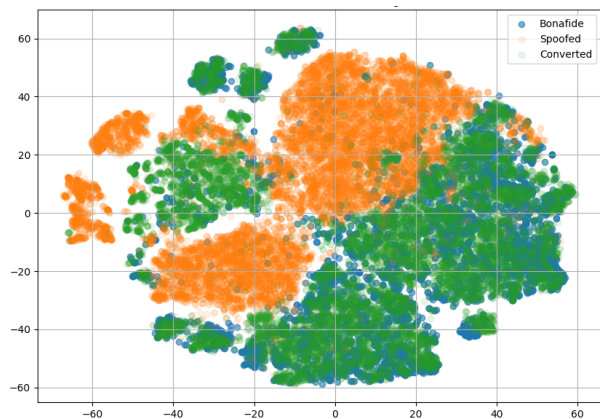


Fig. 2. 合法嵌入来自 LibriSpeech。

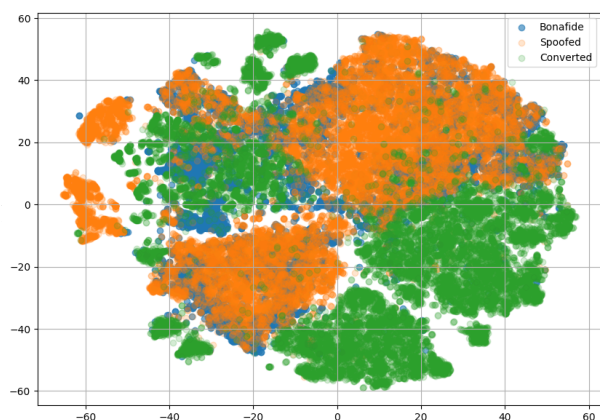


Fig. 3. 合法嵌入来自 ASVspoof2019。

5.2. 声码器的作用

与其它攻击相比 (参见表 2, 第 5 列, 行 A17₅–A31₅), 在 $AASIST_5$ 中 OT_2 的相对较低的等错误率 (表 1) 可能由 vocoder 重叠解释。大多数 $ASVspoof5$ 训练数据是使用 HiFi-GAN vocoder 生成的, 我们的攻击中也使用了该 vocoder。这很可能解释了采用不同声码器的方法 EER 升高的现象, 特别是对于使用 $AASIST_5$ 评估的 $ASVspoof2019$ 方法观察到的极高的 EER (表 2, 第 (5) 列, 行 A07–A19)。

5.3. 最优传输性质

为了进一步说明最优传输的效果, 我们将 t-SNE 应用于 VGGish 嵌入。

图 2 显示了来自 LibriSpeech 正规数据、 $ASVspoof2019$ 生成数据 (A01–A06) 及其 OT 转换版本的嵌入。转换后的嵌入 (Converted) 与 LibriSpeech 目标分布 (Bonafide) 紧密对齐。

图 3 展示了来自 ASVspoof2019 真实数据、ASVspoof2019 生成数据及其 OT 变换对应的数据嵌入。这里，ASVspoof2019 真实数据与 LibriSpeech 之间的差异变得明显。这解释了为什么真实训练数据与 OT₂ 之间的 Fréchet 音频距离 (FAD) 相对较大 (表 3)，以及为何 AASIST₂₀₁₉ 在微调后能够轻易检测到 OT₂，尽管其预微调等错误率较高。

6. 致谢

第一作者感谢张优 (尼尔) 指出了 ASVspoof5 数据集中的对抗性攻击，并将论文 [5] 带到我的注意范围内。

7. REFERENCES

- [1] S. Chen, Ch. Wang, Zh. Chen, et al., “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [2] M. Baas, B. van Niekerk, and H. Kamper, “Voice conversion with just nearest neighbors,” in *Proc. Interspeech*, Apr. 2023, vol. II, pp. 803–806.
- [3] A. Asadulaev, Korst R., et al., “Optimal transport maps are good voice converters,” in *arXiv preprint arXiv:2411.02402*, 2024.
- [4] Selitskiy A. and Kocharekar M., “Discrete optimal transport and voice conversion,” in *arXiv preprint arXiv:2505.04382*, 2025.
- [5] R. Zhang, J. Wei, et al., “SHDA: sinkhorn domain attention for cross-domain audio anti-spoofing,” *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 6474–6489, 2024.
- [6] M. Panariello, W. Ge, H. Tak, M. Todisco, and N. Evans, “Malafide: a novel adversarial convolutive noise attack against deepfake and spoofing detection systems,” in *Proc. Interspeech*, Apr. 2023, pp. 2868–2872.
- [7] X. Wang, H. Delgado, et al., “ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech,” *Computer Speech & Language*, vol. 95, pp. 1–27, 2026.
- [8] Xin Wang, Junichi Yamagishi, et al., “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” *Computer Speech & Language*, vol. 64, pp. 101–114, 2020.
- [9] J. Jung, H.-S. Heo, et al., “AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” in *arXiv preprint arXiv:2110.01200*, 2021.
- [10] J. Yamagishi, C. Veaux, and K. MacDonald, “CSTR VCTK Corpus: english multi-speaker corpus for CSTR voice cloning toolkit (version 0.92),” 2019.
- [11] V. Panayotov, G. Chen, et al., “LibriSpeech: An ASR corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.
- [12] LibriSpeech Clean, ,” <https://www.kaggle.com/datasets/victorling/librispeech-clean>.
- [13] ASVspoof 2019, ,” <https://www.kaggle.com/datasets/awsaf49/asvspoof-2019-dataset>.
- [14] R. Flamary, N. Courty, et al., “POT: Python optimal transport,” *Journal of Machine Learning Research*, vol. 22, no. 78, pp. 1–8, 2021.
- [15] N. Brümmer and J. du Preez, “Application-independent evaluation of speaker detection,” *Computer Speech & Language*, vol. 20, pp. 230–275, 2006.

- [16] A. Gui, H. Gamper, Braun S., and D. Emmanouilidou, “Adapting Fréchet audio distance for generative music evaluation,” in Proc. ICASSP, 2024.
- [17] Sh. Hershey, S. Chaudhuri, et al., “CNN architectures for large-scale audio classification,” in Proc. ICASSP, 2017.
- [18] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, “Intriguing properties of neural networks,” in International Conference on Learning Representations (ICLR), 2014.
- [19] I. Goodfellow, J. Shlens, and Ch. Szegedy, “Explaining and harnessing adversarial examples,” in International Conference on Learning Representations (ICLR), 2015.
- [20] N. Papernot, P. McDaniel, and I. Goodfellow, “Transferability in machine learning: From phenomena to black-box attacks using adversarial samples,” arXiv preprint arXiv:1605.07277, 2016.
- [21] M. Alzantot, B. Balaji, M. Srivastava, et al., “Did you hear that? adversarial examples against automatic speech recognition,” arXiv preprint arXiv:1801.00554, 2018.
- [22] X. Liu, X. Wang, M. Sahidullah, et al., “Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 31, pp. 2507–2522, 2023.

Table 2. EER↓ 在微调之前 (2019 和 5); 经过 OT₂ 攻击之后 (2019_{OT} 和 5_{OT}); 经过攻击和微调之后 (2019_{OT^{FT}} 和 5_{OT^{FT}})。

Attack	2019	2019 _{OT}	2019 _{OT^{FT}}	5	5 _{OT}	5 _{OT^{FT}}
A07	0.52	0.51	0.12	33.28	4.39	10.05
A08	0.42	3.70	0.11	19.90	4.23	9.49
A09	0.00	0.69	0.05	7.79	2.92	7.44
A10	0.86	0.73	0.13	39.42	6.39	14.30
A11	0.18	0.71	0.13	5.65	5.40	13.71
A12	0.78	0.73	0.09	55.41	7.40	13.04
A13	0.15	0.51	0.06	65.51	6.38	15.07
A14	0.15	0.81	0.05	15.34	4.34	8.15
A15	0.55	0.77	0.05	7.30	3.94	8.88
A16	0.65	1.98	0.11	71.97	11.37	17.15
A17	1.26	13.34	0.37	73.26	17.69	12.36
A18	2.61	12.26	0.27	77.73	24.19	14.74
A19	0.65	12.98	0.27	72.57	19.02	14.07
A17 ₅	1.36	-	-	11.43	-	-
A18 ₅	0.43	-	-	51.92	-	-
A19 ₅	0.18	-	-	57.93	-	-
A20 ₅	0.11	-	-	49.78	-	-
A21 ₅	0.72	-	-	13.28	-	-
A22 ₅	1.24	-	-	14.07	-	-
A23 ₅	0.15	-	-	28.81	-	-
A24 ₅	1.39	-	-	10.69	-	-
A25 ₅	0.35	-	-	22.29	-	-
A26 ₅	1.86	-	-	27.41	-	-
A27 ₅	0.18	-	-	24.10	-	-
A28 ₅	0.94	-	-	23.57	-	-
A29 ₅	0.58	-	-	6.83	-	-
A30 ₅	0.25	-	-	39.89	-	-
A31 ₅	0.18	-	-	26.53	-	-

Table 3. FAD↓ 在 (真实, 欺骗), (真实, OT) 和 (欺骗, OT) 之间。

BF Dataset	BF-Spoof	BF-OT	Spoof-OT
LibriSpeech	4.742	0.508	3.665
ASVspoof2019	1.289	3.402	3.665