IO-RAE：用于音频隐私保护的信息混淆可逆对抗样本

Jiajie Zhu[1], **Xia Du**[1*], **Xiaoyuan Liu**[1], **Jizhe Zhou**[2], **Qizhen Xu**[1], **Zheng Lin**[3], **Chi-Man Pun**[4]

[1]厦门理工学院计算机与信息工程学院，中国厦门 [2]四川大学计算机科学学院、机器学习与行业智能工程研究中心，中国成都[3]香港大学电气与电子工程系，中国香港特别行政区薄扶林[4]澳门大学计算机与信息科学系，中国澳门 cosmics36@163.com, duxia@xmut.edu.cn, 2322071027@stu.xmut.edu.cn, jzzhou@scu.edu.cn, qzxu@xmut.edu.cn, linzheng@eee.hku.hk, cmpun@umac.mo

arXiv:2601.01239v1 [cs.SD] 3 Jan 2026

摘要

人工智能的快速发展显著推动了语音识别技术的应用，促使其广泛融入各种应用场景。然而，使用量的激增也凸显出一个关键问题：音频数据极易遭受未经授权的暴露和分析，从而给企业和个人带来严重的隐私风险。本文提出了信息混淆可逆对抗样本（IO-RAE）框架，这是首个旨在利用可逆对抗样本保护音频隐私的方法。IO-RAE 利用大语言模型生成具有误导性但符合语境的内容，有效防止人类和自动语音识别（自动语音识别）系统的未经授权窃听。此外，我们提出了累积信号攻击技术，通过针对低频信号来减轻高频噪声并提升攻击效能。我们的方法确保在不降低音频质量或可用性的情况下保护音频数据。实验评估表明，我们的方法具有优越性：在包括谷歌商业黑盒系统在内的多个自动语音识别模型上对目标关键词进行混淆时，定向误导率达到 96.5%，非定向误导率更达到 100%。此外，通过语音质量感知评估得出的恢复音频质量评分达到 4.45，与高质量原始录音相当。值得注意的是，经自动语音识别系统处理的恢复音频错误率为 0%，表明其几乎实现了无损恢复。这些结果凸显了 IO-RAE 框架在保护敏感音频隐私方面的实际适用性和有效性。

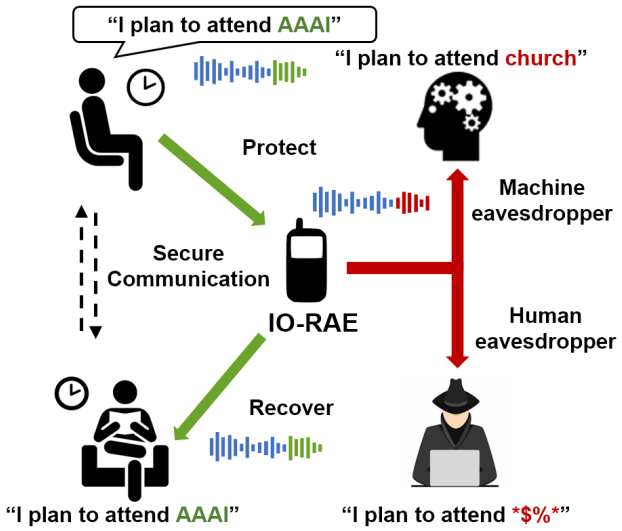

图 1：IO-RAE 通过对抗噪声挫败人类和机器窃听者，并能在获得授权后无损恢复原始音频，从而确保通信安全。

引言

深度神经网络（DNNs）凭借在各类任务中的卓越表现，彻底改变了众多领域（Brown 等，2020；Fang 等，2026；Dosovitskiy 等，2020；Lin 等，2024c；Duan 等，2025b；Lin 等，2025c；Yuan 等，2025b；Lin 等，2025a；Fang 等，2024；Lin 等，2024a；Zhang 等，2025a；Lin 等，2024b；Zhang 等，2024；Yuan 等，2025a；Duan 等，2025a；Lin 等，2025b）。其中，自动语音识别（ASR）已成为一项关键应用，并广泛融入日常生活的方方面面（Han 等，2020；Sainath 等，2021；Ng 等，2021）。然而，自动语音识别技术的迅速且缺乏监管的发展，引发了人们对隐私和安全的重大担忧。恶意商业实体日益利用私人语音录音谋取利益，这凸显了保护音频数据完整性的迫切需求。

近期研究利用数据匿名化技术，通过去除身份信息或扭曲语音信号中的个人特征来保护用户隐私，防止自动语音识别系统提取敏感信息。Chou 等人

*XiaDu 为通讯作者。版权所有 © 2026，人工智能促进协会（www.aaai.org）。保留所有权利。

等人（Chou 等人，2018年）采用对抗学习进行多目标语音转换，无需平行数据即可有效去除说话人身份，从而便于实现匿名化。Champion等人。（Champion，2023年）开发了一种框架，通过修改声学特征和嵌入来分离说话人特有信息。Feng等人。（Feng 等人，2022）提出了域自适应噪声注入方法，在保护隐私的同时降低身份泄露风险。尽管这些方法代表了隐私保护领域的重大进展，但仍面临挑战。恶意攻击者仍可利用先进的深度神经网络提取敏感信息，而匿名化过程往往会降低语音清晰度和可懂度，限制其实际应用。

鉴于现有匿名化方法的局限性，我们需要更有效的方法来平衡隐私保护与语音质量。本文将可逆对抗样本（RAE）（Liu 等人，2023）应用于音频隐私保护：RAE 将对抗攻击与可逆数据隐藏（RDH）相结合，使精心构造的扰动能够误导恶意深度神经网络，之后再由授权用户移除，从而完全恢复原始音频。然而，当前的 RAE 设计面临两大挑战：（1）要求扰动不可感知会限制其混淆能力，导致敏感信息仍可被人类听众获取（Yang 等人，2021；Zhao、Dua 和 Singh，2017；Elsayed 等人，2018）；（2）可逆数据隐藏容量限制要求严格控制嵌入数据的大小，而对于自动语音识别而言，创建覆盖关键内容的对抗补丁尤其困难，因为在音频中定位关键时间戳并非易事（Chen 等人，2024）。

为解决这些问题，我们提出了信息混淆可逆对抗样例（IO-RAE）框架（图 1）。IO-RAE 利用对齐技术定位敏感时间戳，并对这些片段实施定向语音衰减和对抗补丁，从而在保留整体可懂度的同时提升混淆效果。我们进一步采用大语言模型（LLM）（Radford 等人 2018年；Bai 等人 2025年）为补丁生成欺骗性标签，使自动语音识别转录文本听起来连贯可信；此外，还提出累积信号攻击（CSA），这是一种通过将信号能量累积至低频区域来降低高频刺耳感的扰动优化方法。

总而言之，本文的主要贡献如下：

- 我们提出了用于音频隐私保护的 IO-RAE，旨在防止人类和机器窃听者从受保护音频中提取关键信息。据我们所知，这是 RAE 首次应用于音频领域。

- 我们提出了一种基于大语言模型的目标生成方法，以取代传统的预设目标，并利用贪心算法找出最优目标。该方法在显著提升攻击效果的同时，保留了句子的连贯性与合理性。
- 我们提出了用于噪声预处理的 CSA 方法。CSA 通过抑制高频信号来平滑剧烈的扰动，同时通过增强低频信号来维持攻击效果。

背景

对抗攻击

对抗攻击对深度学习模型构成了根本性威胁，其脆弱性已在多模态任务中得到验证。Szegedy 等人（Szegedy 等人，2013）首次揭示，不可感知扰动能够显著改变图像分类模型的预测结果。Goodfellow 等人（Goodfellow、Shlens 和 Szegedy，2014）随后提出了快速梯度符号法（FGSM），这是一种简单而有效的单步攻击，其定义如下：

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x \ell(x, y)), \tag{1}$$

其中，x 为输入，y 为真实标签，ϵ 为扰动幅度。此后，许多攻击方法（Zhao、Liu 和 Larson，2021年；Xiao 等，2019年；Dai 等，2018年）已扩展至图像任务以外的领域。

在音频领域，Carlini 和 Wagner（Carlini 和 Wagner，2018年）率先提出了定向攻击，通过最小化 CTC 损失（Graves 等，2006年）并限制扰动幅度，使对抗语音能够被转录为任意短语。随后，Yakura 等 （Yakura 和 Sakuma，2018年）对室内声学和噪声等环境因素进行建模，以实现稳健的空中攻击。Qin等（Qin 等，2019年）提出了一种考虑心理声学的框架，利用听觉掩蔽生成难以察觉但有效的对抗音频，并将人类听觉模型融入攻击过程。

可逆对抗攻击

可逆对抗攻击的研究（Zhang 等，2025c，b）主要聚焦于图像数据保护，而其在音频数据领域的应用仍未得到充分探索。Liu等人（Liu 等人，2023）首次提出了可逆对抗攻击的概念，将可逆数据隐藏技术与对抗样本相结合，开创了这种新型攻击方法。随后，Xiong 等人（Xiong 等人，2023）将这项工作拓展至黑盒场景，并结合集成建模技术，展示了可逆对抗样本在不同模型中的应用潜力。与此同时，Zhang 等人（Zhang 等人，2022）提出了一种基于 RGAN 的方法，以此取代可逆数据隐藏技术。具体而言，他们通过攻击编码器网络高效生成对抗样本，并利用恢复解码器网络有效还原这些样本，从而实现图像的高效恢复。尽管该方法在恢复图像原始状态方面表现出色，但在针对黑盒模型的对抗任务中效果有限。最近，DP-RAE 框架（Zhu 等人，2024）通过双阶段设计提高了黑盒场景中的攻击成功率。本文提出一种将 LSB（Kavitha、Koshti 和 Dunghav，2012年）与对抗样本相结合的方法，以有效保护音频数据。

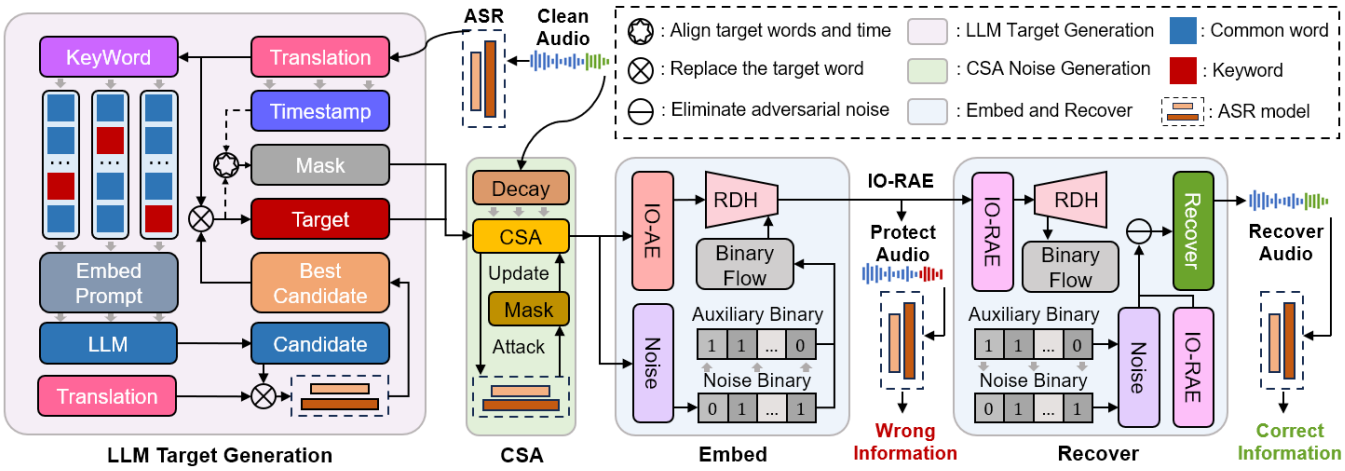

图2：IO-RAE 框架概览。

所提方法

概述

本节介绍 IO-RAE 框架，该框架通过 CSA 生成对抗扰动，以混淆关键信息。如图 2 所示，IO-RAE 由三个部分组成：LLM目标生成、通过 CSA 创建对抗扰动，以及可逆扰动嵌入与恢复。首先，IO-RAE 使用 KeyBERT（Grootendorst 2020）从正确转录文本中提取关键信息。随后，将这些关键词与提示词结合并输入大语言模型，以生成逻辑连贯的替换词。接着，使用蒙特利尔强制对齐器 MFA（McAuliffe等，2017）技术对齐文本和音频，获取对应的掩码。对这些掩码区域进行音频衰减和扰动生成，并使用 CSA 增强低频成分。最后，将辅助信息和扰动矩阵嵌入对抗音频中，形成 IO-RAE。在恢复过程中，可逆数据隐藏会提取并解码扰动矩阵，从而有效去除对抗扰动。

LLM目标生成

与基于固定短语或通过最大化词向量距离来生成目标短语的方法不同，我们利用大语言模型生成候选替换词。这种方法避免了向量距离最远的词所造成的混淆，并提升了生成短语的多样性。我们将目标转录文本表示为 $y = \{y_0, y_1, ..., y_n\}$，其中 y 最初设为正确的转录文本，而 y_i 表示转录文本中的 i_{th} 词。为实现断章取义的目标，我们需要找出一个子集 $y^* \subseteq y$，用来替换 y 并有效扭曲原意。因此，我们使用 KeyBERT 找出 y 中影响力最大的词，以推导出 y^*，并利用大语言模型生成候选词。这样便生成了与 y^*相关的候选词。此外，图像识别任务中误导不同类别的难度各不相同，自动语音识别系统中也存在这一现象。这种差异取决于输入与目标 y之间的损失函数 ℓ 的值：

$$\ell(x, \delta, y) = \ell_{CTC}(f(x+\delta), y), \tag{2}$$

其中 δ 表示对抗噪声，x 表示原始音频。因此，我们可以得出结论：在众多候选词中，存在一个最优替换集 y_*，能够以最低成本误导自动语音识别系统。为高效确定 y^*，我们采用贪心算法，为 y^*中的每个位置找出最佳替换词。随后，我们利用 MFA 提供的各词时间戳，为扰动 δ生成位置掩码 $\mathbb{M}$。整体算法详见算法1。

算法1：LLM目标生成

输入：原始音频 x。输出：输出：目标转录文本 y，位置掩码 $\mathbb{M}$。初始化：$y = f(x)$，$y^* = \{\}$，$\mathcal{L} = \{\}$，$\mathbb{M} \leftarrow 0\,[shape(x)]$1：获取关键词：$y^* = \text{KeyBERT}(y)$；2：使用 MFA 对齐 x 和 y；3：对于 $i = 0 \rightarrow len(y^*) - 1$ 执行 4：将 Embedding y_i^* 嵌入提示词；5：获取候选词库 cl：cl = 大语言模型(提示词)；6：对于 $j = 0 \rightarrow len(cl) - 1$ 执行 7：用 cl_j替换$y_{\,idx(y_i^* \text{ in } y)}$；8：使用公式 2计算损失 ℓ；9：将 ℓ追加到 $\mathcal{L}$；10：结束循环 11：找到 $\mathcal{L}$中值最小项的索引 min；12：用 cl_{min}替换$y_{\,idx(y_i^* \text{ in } y)}$；13：获取$y_{\,idx(y_i^* \text{ in } y)}$的起始时间和结束时间；14：$\mathbb{M}\,[start : \ end] = 1$；15：结束循环 16：返回 y，$\mathbb{M}$

利用大语言模型和基于贪心的概念找到的最优目标转录文本 y 和位置掩码$\mathbb{M}$ 将

用于后续的 CSA攻击，以生成目标扰动。

累积信号攻击

音频数据存储具有独特性：由于采样率高，即使是较短的音频片段也需要大量张量来表示，因此完整扰动矩阵的数据量十分庞大。鉴于可逆数据隐藏技术的存储容量有限，嵌入前必须压缩这些扰动。因此，在 CSA 中，我们采用复合采样来减小数据量，同时保持攻击效果。此外，CSA 通过避免为已成功攻击的词语生成额外扰动，并使用累积信号减弱刺耳噪声，来抑制音频攻击中的过拟合。

CSA 方法首先衰减原始音频内容，以确保受保护材料不可感知；随后对矩阵执行复合采样操作，确保压缩梯度矩阵保留其对抗属性。假设复合样本的大小为 s，压缩矩阵为 $\mathcal{L}_c$。$\mathcal{L}_c$ 中 i_{th} 元素的值可表示为：

$$\mathcal{L}_c[i] = \sum_{j=i*s}^{i*s+s} \nabla_x \ell(x, \delta, y)_j / s, \tag{3}$$

然后，CSA 方法将区间 $[i: i+s]$ 内的原始梯度替换为平均压缩梯度$\mathcal{L}_c[i]$。这一调整使扰动能够与压缩梯度保持一致。

随后，CSA 方法通过增强低频分量来优化生成的扰动。经过离散傅里叶变换（Cooley和Tukey（1965））后，扰动信号 δ 在频域中表示为：

$$\triangle(k) = \sum_{n=0}^{N-1} \delta[n] \, e^{-j\frac{2\pi kn}{N}}, \tag{4}$$

累积信号 c 对应于 δ 的频域表示为：

$$C(k) = \sum_{m=0}^{N-1} \delta[m] \left(\sum_{n=m}^{N-1} e^{-j\frac{2\pi kn}{N}} \right), \tag{5}$$

对于每个固定的 m，可得：

$$\sum_{n=m}^{N-1} e^{-j\frac{2\pi kn}{N}} = e^{-mj\frac{2\pi kn}{N}} \sum_{n=0}^{N-1-m} e^{-j\frac{2\pi kn}{N}}. \tag{6}$$

当 N 足够大时，公式 6 可近似为：

$$\sum_{n=0}^{N-1-m} e^{-j\frac{2\pi kn}{N}} \approx \frac{1}{1 - e^{-j\frac{2\pi k}{N}}}, \tag{7}$$

$C(k)$ 与 $\triangle(k)$ 之间的关系可表示为：

$$C(k) \approx \triangle(k) \cdot \frac{1}{1 - e^{-j\frac{2\pi k}{N}}}. \tag{8}$$

因此，对受扰信号进行累积操作会增强低频分量，同时抑制高频分量。详细推导见补充材料。

为防止扰动生成过程中对特定词语过拟合，CSA 方法会在每次迭代前评估每个目标替换词的攻击成功率，并更新成功矩阵 $\mathbb{S}$。如果某个词已成功受到攻击，则后续迭代中不再对其施加额外扰动。完整算法见算法 2。

算法2：累积信号攻击

输入： 原始音频 x，目标 y，衰减率 α，掩码 $\mathbb{M}$，迭代 I，单元噪声 ϵ，乘法因子 t
输出： 对抗音频 x_{adv}，噪声矩阵 δ
初始化： $g_0 = 0$，$x^0_{adv} = \frac{1}{\alpha} x \odot \mathbb{M}$，CSA噪声 $\eta = 0$，$\delta \leftarrow 0$[形状(x)], $\eta \leftarrow 0$[形状(x)], $\mathbb{S} \leftarrow \mathbb{M}$, $\epsilon = \epsilon \cdot \frac{1+\alpha/5}{100}$

对于 $i = 0 \to I-1$ **执行**
  $y_{adv} = f\left(x^i_{adv} + \eta\right)$
  验证攻击是否成功于 y_{adv}
  根据 y_{adv} 的结果更新 $\mathbb{S}$
  计算公式 2 的损失并得到 g_{t+1}
  $g_{t+1} = \mu \cdot g_t + \frac{g_{t+1}}{\|g_{t+1}\|_2}$
  使用公式 3 压缩并更新 g_{t+1}
  $\delta = \delta + \epsilon \cdot$ 符号($\mathbb{M} \odot \mathbb{S} \cdot g_{t+1}$)
  $\delta =$片段$^{t\cdot\delta}_{\delta}$ (δ)
  $\eta = \frac{1}{\alpha} \cdot \sum_{m=0}^{N-1} \delta[m]$ ▷ 累积信号
  $x^{i+1}_{adv} = x^i_{adv} + \eta$
结束循环
返回 x_{adv}, δ

随后，在下一阶段使用对抗样本及其对应的扰动矩阵来构建可逆对抗音频攻击。值得注意的是，为了便于信息编码并减少存储开销，返回的扰动矩阵对应于累积信号操作之前的原始矩阵 δ。

嵌入与恢复

在 IO-RAE 框架中，我们采用可逆数据隐藏技术，将附加数据嵌入对抗样本。首先，对 CSA 方法生成的扰动矩阵进行编码。具体而言，以 s 为间隔对扰动进行采样，提取标准扰动 ϵ 的系数，i_{th} 元素的消息可表示为：

$$Mes_i = \frac{\sum_{j=i*s}^{i*s+s} \delta_j}{\epsilon \times s}. \tag{9}$$

依次遍历所有有效的扰动矩阵，并汇总其对应的 Mes_i，形成 Mes，从而得到扰动数据流。随后，将该数据流与辅助信息结合，包括扰动的起始位置、标准扰动幅度和衰减率。最终，将组合后的信息嵌入对抗音频中，以确保其对抗性得以保留。

在音频恢复阶段，首先使用可逆数据隐藏技术提取嵌入信息。利用提取出的 Mes 和辅助数据，重建扰动矩阵及

其对应的嵌入位置。随后，将 CSA 的正向过程重新应用于扰动矩阵，以获得 CSA噪声 η。最后，通过逐步去除对抗音频中的扰动，有效恢复原始音频。

实验

实验设置

数据集。 我们从 Mozilla Common Voice（Ardila 等人，2019年）、TIMIT（Garofolo 等人，1993年）和 LibriSpeech（Panayotov 等人，2015年）中分别随机选取了 200 条经过准确转录的音频样本，以验证该方法在不同数据来源上的有效性。

环境。鉴于 DeepSpeechV3 在语音识别领域的重要性，我们采用广泛使用的 DeepSpeechV3（Battenberg 等人，2017年）作为主要目标模型。为评估可迁移性，我们还在 Conformer（Gulati 等人，2020年）、Whisper（Radford 等人，2023年）、Wenet（Yao 等人，2021年）、SenseVoice（Gao 等人，2023年）和 Wav2vec（Baevski 等人，2020年）上进行了测试。所有实验均使用 PyTorch 实现，并在 NVIDIA A40 GPU 上运行。

评估指标。 为了评估音频质量和可懂度，我们使用了信噪比、尺度不变信号失真比、短时客观可懂度和感知语音质量评估。为了评估攻击效果，我们评估了定向攻击成功率（TSR）、定向词错误率（TWER）、非定向攻击成功率（USR）和非定向词错误率（UWER）。

参数设置。 我们将单元噪声 ϵ 设为 $1/5$，将乘数因子 t 设为 15，这意味着每个扰动仅需 5 比特。复合样本大小 s 设为 2，以平衡存储与效果。音频衰减 α 设为信号最大值每 0.1 衰减 5，上限为 30。攻击迭代次数 I 固定为 200。对于目标词生成，我们采用了 Qwen2.5-VL-7B。

攻击与恢复能力

本节展示 IO-RAE 在三个主要数据集 Common、TIMIT 和 LibriSpeech 上针对 DeepSpeechV3 模型进行攻击与恢复的性能。评估维度主要包括音频质量和攻击性能。

关于攻击性能，如表 1 所示，IO-RAE 在多个数据集上展现出较强的目标误导能力。值得注意的是，在 LibriSpeech 数据集上，TSR 达到 96.5%，而 TWER 仅为 0.21%，这表明即使目标攻击未完全成功，转录结果与目标输出之间的偏差仍然很小。因此，自动语音识别系统难以从对抗音频中提取准确信息。如图 3 所示，在梅尔频谱图（Stevens、Volkmann 和 Newman，1937）中，遭受攻击的音频区域通过 CSA 得到了全面保护。

在恢复方面，图 3 直观地展示了恢复音频不含误导性扰动，并完美恢复至干净样本状态。表 1 中的实验结果进一步证实，恢复音频在各项测试中均表现出卓越的质量，其感知语音质量评估（PESQ）得分为 4.45，接近 4.5 的理论最大值，客观验证了恢复音频的优越性。其他关键音频质量指标，如信噪比（SNR）、尺度不变信号失真比（SISDR）和短时客观可懂度（STOI），也均取得了显著提升。使用 DeepSpeechV3 对恢复音频进行重新转录后，目标成功率（TSR）达到 0%，有效恢复了原始信息，证明了我们方法的有效性和实用性。

来源: Common	音频质量 SNR	尺度不变信号失真比	STOI	PESQ	攻击性能 TSR	TWER
IO-AE	1.8	-0.22	0.83	2.6	96.0	0.6
IO-RAE	1.8	-0.22	0.83	2.6	96.5	0.55
恢复	53.06	53.06	0.99	4.38	0	16.18
来源: TIMIT	音频质量 SNR	尺度不变信号失真比	STOI	PESQ	攻击性能 TSR	TWER
IO-AE	-12.37	-15.35	0.78	2.41	96.0	1.08
IO-RAE	-12.37	-15.35	0.78	2.41	95.0	1.36
恢复	42.76	42.76	0.99	4.28	0	17.31
来源: LibriSpeech	音频质量 SNR	尺度不变信号失真比	STOI	PESQ	攻击性能 TSR	TWER
IO-AE	3.37	1.99	0.89	3.02	96.5	0.21
IO-RAE	3.28	1.91	0.89	3.02	96.5	0.21
恢复	54.75	54.75	0.99	4.45	0	10.2

表 1：对抗音频（IO-AE 和 IO-RAE）及恢复音频的信噪比（dB）、尺度不变信号失真比（dB）、短时客观可懂度、感知语音质量评估、目标成功率（%）和目标词错误率（%）。恢复结果以红色突出显示。

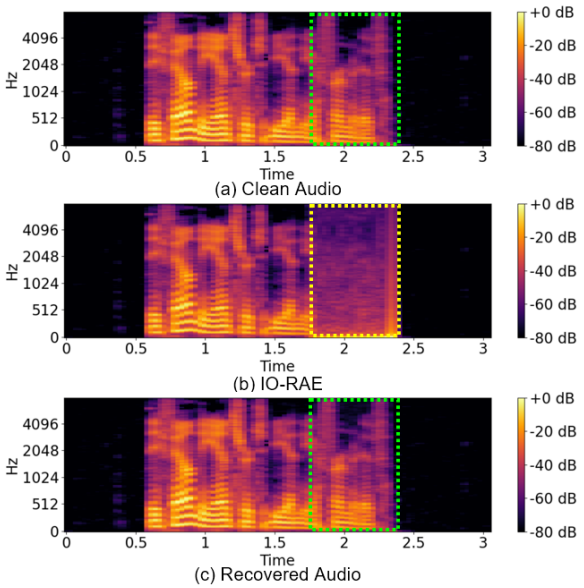

图 3：（a）、（b）和（c）分别表示干净音频、IO-RAE 和恢复音频的梅尔频谱图。

鲁棒性评估

在真实场景中，输入音频通常会经过一系列防御方法，以减轻对抗噪声对模型的影响。因此，确保对抗样本的鲁棒性至关重要。为了评估

来源：Common	防御方法						
	AS	AT	LPF	MS	QT	MP3-V	OPUS
目标成功率 ↑	8.0	45.0	26.0	13.5	**59.0**	25.0	19.5
目标词错误率 ↓	22.7	9.45	13.79	24.17	**6.89**	13.39	15.99
USR ↑	**100**	**100**	**100**	**100**	98.5	**100**	**100**
UWER ↑	24.13	18.33	19.29	**29.19**	17.14	18.31	19.86
来源：TIMIT	防御方法						
	AS	AT	LPF	MS	QT	MP3-V	OPUS
目标成功率 ↑	4.5	32.0	38.0	20.5	34.5	19.5	9.0
目标词错误率 ↓	27.27	18.09	11.72	23.82	13.17	16.63	19.28
USR ↑	99.5	99.5	**100**	**100**	**100**	99.0	99.5
UWER ↑	28.98	25.75	20.45	**30.31**	23.31	21.66	22.27
来源：LibriSpeech	防御方法						
	AS	AT	LPF	MS	QT	MP3-V	OPUS
目标成功率 ↑	10.5	57.5	50.5	12.5	**68.5**	32.0	30.0
目标词错误率 ↓	12.69	4.58	5.11	13.9	**3.1**	8.13	8.07
USR ↑	**99.5**	**99.5**	**99.5**	**99.5**	99.0	**99.5**	**99.5**
UWER ↑	14.01	10.69	11.03	**16.91**	10.43	11.77	11.38

表 2：各种防御方法下 IO-RAE 的 TSR（%）、TWER（%）、USR（%）和 UWER（%）。

IO-RAE 的鲁棒性，我们对对抗样本应用了多种防御方法，包括平均平滑（AS）（Du 等人，2020）、音频湍流（AT）（Yuan 等人，2018）、低通滤波器（LPF）（Kwon、Yoon 和 Park，2019）、中值平滑（MS）（Yang 等人，2018）、量化（QT）（Yang 等人，2018）、MP3-V 和 OPUS（Vos 等人，2013）。

如表 2 所示，尽管这些防御方法对定向攻击的影响程度各不相同，但 IO-RAE 生成的对抗样本在所有数据集上的误导率始终接近 100%，几乎达到完美水平，有效阻碍了对原始信息的准确解读。这种鲁棒性可归因于 IO-RAE 的可逆对抗攻击特性，该特性能够施加更强的扰动，同时不损害整体音频质量。因此，IO-RAE 在实际应用中展现出强大的鲁棒性和可靠性。

可迁移性评估

由于入侵者所使用模型的具体架构通常未知，因此有必要验证 IO-RAE 是否能够在不同模型间保护关键目标信息。在图像攻击领域，由于模型架构具有相似性，可以通过计算集成梯度来生成可迁移的扰动，从而有效攻击未知模型。然而，在音频攻击领域，模型架构之间存在显著差异，使得跨模型迁移更加困难和具有挑战性。IO-RAE 利用 CSA 有效隐藏关键信息，实现跨模型保护。我们选取了若干架构差异很大的模型来评估 IO-RAE 的可迁移性，包括 Wenet、Conformer、SenseVoice、Whisper 和 Wav2vec。

如表 3 所示，IO-RAE 能够有效保护不同模型中的关键信息，阻止多个模型获取正确的原始信息。这表明 IO-RAE 具有很强的跨模型信息安全能力，即使面对不同架构，也能持续保护关键信息。

来源：模型	Common		TIMIT		ibrispeech	
	USR ↑	UWER ↑	USR ↑	UWER ↑	USR ↑	UWER ↑
Wenet	91.5	18.25	89.5	17.48	85.5	10.63
Conformer	94.0	19.72	98.0	21.61	92.0	11.38
SenseVoice	96.0	21.06	98.0	19.85	93.5	13.23
Whisper	99.5	**29.01**	**100**	**29.59**	**99.5**	**29.26**
Wav2vec	**100**	22.67	**100**	22.62	99.0	12.22

表 3：针对不同自动语音识别系统的 IO-RAE 的 USR（%）和 UWER（%），该 IO-RAE 由 DeepSpeechV3 生成。

商业模型攻击

为进一步验证我们的模型在真实场景中的有效性，我们针对商业自动语音识别系统谷歌 API [1]开展了攻击实验。我们的目标是利用 IO-RAE 欺骗谷歌 API，使其无法从音频信号中准确提取正确信息。我们从 Common Voice 数据集中准备了 50 个语音样本，API 最初能够准确转录所有样本。随后，我们生成了相应的 IO-RAE 来欺骗该 API。得益于 CSA 引入的干扰，IO-RAE 完全保护了关键信息，实现了 100% 的攻击成功率。此外，我们还从 IO-RAE 中恢复了音频，恢复音频再次能够被谷歌 API 正确识别，这意味着恢复音频消除了对抗性扰动并恢复了音频质量。我们将相应的音频数据上传至匿名网站 [2]，以便播放和下载，从而试听攻击与恢复效果。

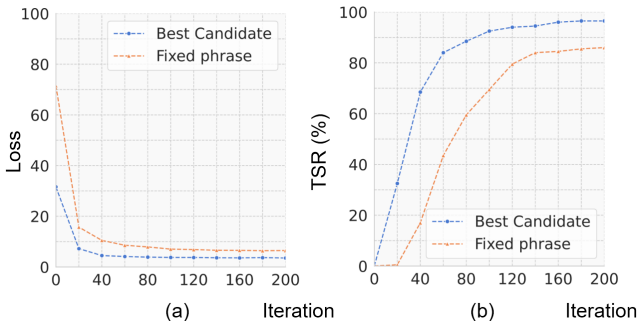

图 4：（a）表示 Common 数据集上两种策略的 CTC 损失，（b）表示目标成功率（%）。

消融研究

我们首先探究了“最佳候选策略”对攻击性能的影响。如图 4 所示，与“固定策略”相比，采用“最佳候选策略”后，攻击性能显著提升。这是因为目标词与输入音频之间的损失越高，自动语音识别感知到的匹配度就越低，

[1]https://cloud.google.com/speech-to-text [2] https://sites.google.com/view/io-rae/io-rae/

来源：Common	复合样本								单元噪声					
	IO-AE				IO-RAE				IO-AE			IO-RAE		
	$s=2$	$s=3$	$s=4$	$s=5$	$s=2$	$s=3$	$s=4$	$s=5$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$
目标成功率 ↑	96.0	92.0	91.0	84.5	96.5	93.0	89.5	84.5	96.0	96.0	93.0	95.5	96.5	92.5
目标词错误率 ↓	0.60	1.42	1.54	3.05	0.55	1.67	1.96	3.16	0.62	0.60	0.60	0.75	0.55	1.33
USR ↑	100	100	100	100	100	100	100	100	100	100	100	100	100	100
UWER ↑	16.21	16.2	16.27	16.8	16.16	16.33	16.27	16.8	16.08	16.21	16.21	16.08	16.16	16.22
来源：TIMIT	复合样本								单元噪声					
	IO-AE				IO-RAE				IO-AE			IO-RAE		
	$s=2$	$s=3$	$s=4$	$s=5$	$s=2$	$s=3$	$s=4$	$s=5$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$
目标成功率 ↑	96.0	91.0	84.5	83.5	95.0	88.5	83.5	82.0	88.5	96.0	93.5	85.5	95.0	90.5
目标词错误率 ↓	1.08	2.03	3.30	3.38	1.36	2.61	3.52	3.62	2.09	1.08	1.55	2.57	1.36	2.02
USR ↑	100	100	100	100	100	100	100	100	100	100	100	100	100	100
UWER ↑	17.74	17.6	18.07	17.92	17.74	17.90	18.1	18.08	17.59	17.74	17.40	17.89	17.74	17.52
来源：LibriSpeech	复合样本								单元噪声					
	IO-AE				IO-RAE				IO-AE			IO-RAE		
	$s=2$	$s=3$	$s=4$	$s=5$	$s=2$	$s=3$	$s=4$	$s=5$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$
目标成功率 ↑	96.5	94.5	90.5	87.5	96.5	94.0	90.5	87.0	94.5	96.5	96.0	94.0	96.5	95.5
目标词错误率 ↓	0.21	0.47	0.95	1.75	0.22	0.52	0.95	1.84	0.57	0.21	0.25	0.71	0.22	0.28
USR ↑	100	100	100	100	100	100	100	100	100	100	100	100	100	100
UWER ↑	10.22	10.34	10.44	10.70	10.22	10.34	10.44	10.70	10.34	10.22	10.23	10.39	10.22	10.26

表4：不同设置下 IO-RAE 的目标成功率（%）、目标词错误率（%）、USR（%）、UWER（%）。复合样本展示了不同复合样本大小下 IO-RAE 的攻击性能，而单位噪声则表示不同强度的对抗噪声对性能的影响。

因此需要更大或更复杂的扰动，才能迫使自动语音识别输出目标词。“最佳候选策略”能有效降低这种损失，从而降低所需的扰动强度并提高目标成功率。

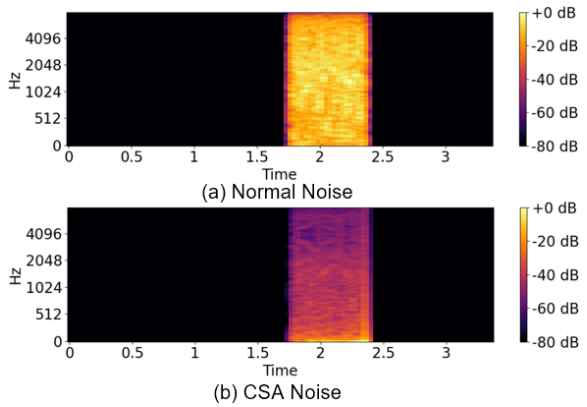

图5：（a）表示普通噪声，（b）表示 CSA噪声（我们的方法）。

为评估 CSA 对尖锐噪声的效果，我们比较了普通噪声和 CSA 生成的噪声。如图5所示，频谱分析表明，普通噪声在中高频仍保留较高能量，而 CSA 显著抑制了这些频段的能量，降低了刺耳感，使噪声频谱更加平滑。

随后，我们考察了两个关键因素：复合样本大小和扰动幅度。如表4所示，较大的复合样本大小会因梯度平均而降低攻击性能，从而削弱扰动方向的精确性。相反，样本过小会因信息密度较高而增加存储需求（表5）。因此，复合样本大小为2时，效率与性能之间达到了平衡。对于扰动幅度，增大尺度始终能提高攻击成功率。我们发现，尺度为 $1/5$ 时，在扰动隐蔽性和攻击效果之间实现了最佳权衡。

	复合样本				
来源	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$
通用	51,603	25,811	17,214	12,915	10,336
TIMIT	47,627	23,823	15,889	11,921	9,541
LibriSpeech	42,296	21,158	14,112	10,589	8,475

表5：不同数据集上不同复合样本大小所需的存储量（比特）。

结论

在本研究中，我们提出了一种用于音频隐私保护的新型可逆对抗攻击框架。IO-RAE 利用大语言模型生成目标攻击短语，并采用累积信号技术无缝融入噪声。IO-RAE 展现出优异的攻击和恢复性能，并已针对谷歌的商用黑盒模型进行测试，在非目标攻击中取得了 100% 的成功率。未来，我们计划探索适用于黑盒场景的高迁移性目标攻击方法，以增强其误导能力，从而实现更实用、更有效的音频数据保护。

致谢

本研究部分得到了厦门市自然科学基金研究项目（3502Z202472028）、厦门市科技计划项目（3502Z20231042）、中央高校基本科研业务费（1082204112364），以及澳门特别行政区科学技术发展基金项目（项目编号：0193/2023/RIA3 和 0079/2025/AFJ）的资助。

参考文献

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2019年。Common Voice：一个大规模多语言语音语料库。*arXiv*预印本 *arXiv:1912.06670*。Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020年。wav2vec 2.0：一种用于语音表示自监督学习的框架。神经信息处理系统进展，33：12449–12460。Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025年。Qwen2. 5-vl技术报告。*arXiv*预印本 *arXiv:2502.13923*。Battenberg, E.; Chen, J.; Child, R.; Coates, A.; Li, Y. G. Y.; Liu, H.; Satheesh, S.; Sriram, A.; and Zhu, Z. 2017年。探索用于端到端语音识别的神经转导器。发表于*2017*年 *IEEE* 自动语音识别与理解研讨会 *(ASRU)*，206–213。IEEE。Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020年。语言模型是少样本学习器。神经信息处理系统进展，33：1877–1901。Carlini, N.; and Wagner, D. 2018年。音频对抗样例：针对语音转文本的定向攻击。发表于*2018*年 *IEEE* 安全与隐私研讨会 *(SPW)*，1–7。IEEE。Champion, P. 2023年。语音匿名化：说话人匿名化技术的评估与设计。*arXiv*预印本 *arXiv:2308.04455*。Chen, Z.; Chai, X.; Gan, Z.; Wang, B.; and Zhang, Y. 2024。RAE-VWP：一种面向医疗物联网的医学图像可逆对抗样例隐私与版权保护方法。*IEEE* 物联网期刊。Chou, J.-c.; Yeh, C.-c.; Lee, H.-y.; and Lee, L.-s. 2018年。无需平行数据，通过对抗学习解耦音频表示实现多目标语音转换。*arXiv*预印本 *arXiv:1804.02812*。Cooley, J. W.; and Tukey, J. W. 1965。复数傅里叶级数机器计算的一种算法。计算数学，19(90)：297–301。Dai, H.; Li, H.; Tian, T.; Huang, X.; Wang, L.; Zhu, J.; and Song, L. 2018年。针对图结构数据的对抗攻击。发表于国际机器学习会议，1115–1124。PMLR。Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020年。一张图像相当于16×16词条：用于大规模图像识别的Transformer。*arXiv*预印本 *arXiv:2010.11929*。Du, T.; Ji, S.; Li, J.; Gu, Q.; Wang, T.; and Beyah, R. 2020. Sirenattack：为端到端声学系统生成对抗音频。载于第*15*届*ACM*亚洲计算机与通信安全会议论文集，357–369。Duan, T.; Zhang, Z.; Guo, S.; Huang, D.; Zhao, Y.; Lin, Z.; Fang, Z.; Luan, D.; Cui, H.; and Cui, Y. 2025a. LEED：面向多智能体强化学习的高效且可扩展的大语言模型赋能的专家示范框架。*arXiv*预印本 *arXiv:2509.14680*。Duan, T.; Zhang, Z.; Lin, Z.; Guo, S.; Guan, X.; Wu, G.; Fang, Z.; Meng, H.; Du, X.; Zhou, J.-Z.; et al. 2025b. 面向移动系统中多智能体强化学习的大语言模型驱动的具备平稳性感知能力的专家示范。*arXiv*预印本 *arXiv:2511.19368*。Elsayed, G.; Shankar, S.; Cheung, B.; Papernot, N.; Kurakin, A.; Goodfellow, I.; and Sohl-Dickstein, J. 2018. 同时欺骗计算机视觉系统和时间受限人类的对抗样本。神经信息处理系统进展，31。Fang, Z.; Lin, Z.; Chen, Z.; Chen, X.; Gao, Y.; and Fang, Y. 2026. 用于大语言模型参数高效微调的自动化联邦流程。*IEEE*移动计算汇刊Fang, Z.; Lin, Z.; Hu, S.; Cao, H.; Deng, Y.; Chen, X.; and Fang, Y. 2024. IC3M：用于监测驾驶员和乘客异常状态的车内多模态多目标监测。*arXiv*预印本 *arXiv:2410.02592*。Feng, T.; Hashemi, H.; Annavaram, M.; and Narayanan, S. S. 2022. 通过域自适应噪声注入增强语音情感识别中的隐私保护。载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*），7702–7706。IEEE。Gao, Z.; Li, Z.; Wang, J.; Luo, H.; Shi, X.; Chen, M.; Li, Y.; Zuo, L.; Du, Z.; Xiao, Z.; and Zhang, S. 2023. FunASR：基础端到端语音识别工具包。载于*INTERSPEECH*。Garofolo, J. S.; Lamel, L. F.; Fisher, W. M.; Fiscus, J. G.; and Pallett, D. S. 1993. DARPA TIMIT声学语音连续语料库CD-ROM。NIST语音光盘1-1.1。*NASA STI/Recon*技术报告*n*，93：27403。Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. 解释并利用对抗样本。*arXiv*预印本 *arXiv:1412.6572*。Graves, A.; Fern´andez, S.; Gomez, F.; and Schmidhuber, J. 2006. 连接时序分类：利用循环神经网络标注未分段的序列数据。载于第*23*届国际机器学习会议论文集，369–376。Grootendorst, M. 2020. KeyBERT：基于BERT的简约关键词提取。Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. 2020. Conformer：用于语音识别的卷积增强型Transformer。*arXiv*预印本 *arXiv:2005.08100*。

Han, W.;Zhang, Z.;Zhang, Y.;Yu, J.;Chiu, C.-C.; Qin, J.; Gulati, A.; Pang, R.; and Wu, Y. 2020. Contextnet：利用全局上下文改进自动语音识别的卷积神经网络。*arXiv*预印本 *arXiv:2005.03191*。

Kavitha,K. K.； Koshti, A.；和Dunghav,P.2012年。使用最低有效位算法进行隐写术。《国际工程研究与应用期刊》，2(3)：338–341。

Kwon, H.; Yoon, H.; and Park, K.-W. 2019年。POSTER：通过音频修改检测音频对抗样本。载于*2019*年*ACM SIGSAC*计算机与通信安全会议论文集，2521–2523。

Lin, Z.; Chen, Z.; Fang, Z.; Chen, X.; Wang, X.; and Gao, Y. 2024年a。Fedsn：异构低地球轨道卫星网络上的联邦学习框架。*IEEE*移动计算汇刊Lin, Z.; Qu, G.; Chen, X.; and Huang, K. 2024年b。6G边缘网络中的分割学习。*IEEE*无线通信Lin, Z.; Qu, G.; Wei, W.; Chen, X.; and Leung, K. K. 2024年c。Adaptsfl：资源受限边缘网络中的自适应分割联邦学习。*IEEE*网络汇刊Lin, Z.; Wei, W.; Chen, Z.; Lam, C.-T.; Chen, X.; Gao, Y.; and Luo, J. 2025年a。分层分割联邦学习：收敛性分析与系统优化。*IEEE*移动计算汇刊Lin, Z.; Zhang, Y.; Chen, Z.; Fang, Z.; Chen, X.; Vepakomma, P.; Ni, W.; Luo, J.; and Gao, Y. 2025年b。HSplit-LoRA：面向大语言模型的异构分割参数高效微调框架。*arXiv*预印本 *arXiv:2505.02795*。Lin, Z.; Zhang, Y.; Chen, Z.; Fang, Z.; Wu, C.; Chen, X.; Gao, Y.; and Luo, J. 2025年c。LEO-Split：低地球轨道卫星网络上的半监督分割学习框架。*IEEE*移动计算汇刊Liu, J.; Zhang, W.; Fukuchi, K.; Akimoto, Y.; and Sakuma, J. 2023年。未经授权的AI无法识别我：可逆对抗样本。《模式识别》，134：109048。

McAuliffe, M.; Socolof, M.; Mihuc, S.; Wagner, M.; and Sonderegger, M. 2017年。蒙特利尔强制对齐器：使用Kaldi进行可训练的文本-语音对齐。载于*Interspeech*，第2017卷，498–502。Ng, E. G.; Chiu, C.-C.; Zhang, Y.; and Chan, W. 2021年。推动非自回归语音识别的极限。*arXiv*预印本 *arXiv:2104.03416*。Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015年。LibriSpeech：基于公有领域有声读物的自动语音识别语料库。载于*2015*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），5206–5210。IEEE。

Qin, Y.; Carlini, N.; Cottrell, G.; Goodfellow, I.; and Raffel, C. 2019年。不可感知、鲁棒且具有目标性的自动语音识别对抗样本。载于国际机器学习会议，5231–5240。PMLR。Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023年。通过大规模弱监督实现鲁棒语音识别。载于国际机器学习会议，28492–28518。PMLR。

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; 等。2018年。通过生成式预训练改进语言模型理解。

Sainath, T. N.； He, Y.； Narayanan, A.； Botros, R.； Pang, R.； Rybach, D.； Allauzen, C.； Variani, E.； Qin, J.； Le-The, Q.-N.； 等。2021年。改进罕见词建模的高效流式处理非循环端侧端到端模型。在*Interspeech*，第8卷，1777–1781。

Stevens, S. S.； Volkmann, J.； 和Newman, E. B. 1937年。用于测量音高心理量值的量表。美国声学学会期刊，8(3)：185–190。

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; 和 Fergus, R. 2013. 神经网络的有趣特性。*arXiv*预印本 *arXiv:1312.6199*。

Vos, K.; Sørensen, K. V.; Jensen, S. S.; 和 Valin, J.-M. 2013. 使用 Opus 进行语音编码。载于音频工程学会第*135*届大会。音频工程学会。

Xiao, C.; Deng, R.; Li, B.; Lee, T.; Edwards, B.; Yi, J.; Song, D.; Liu, M.; 和 Molloy, I. 2019年。Advit：基于视频时间一致性的对抗帧识别器。载于*IEEE/CVF*国际计算机视觉大会论文集，3968–3977。

Xiong, L.; Wu, Y.; Yu, P.; 和 Zheng, Y. 2023年。一种用于共享图像可授权识别的黑盒可逆对抗样本。《模式识别》，140：109549。

Yakura, H.; and Sakuma, J. 2018年. 针对物理攻击的稳健音频对抗样本. *arXiv* 预印本 *arXiv:1810.11793*。

Yang, X.; Dong, Y.; Pang, T.; Su, H.; Zhu, J.; Chen, Y.; a nd Xue, H. 2021年. 通过生成对抗身份掩码实现人脸加密. 收录于 *IEEE/CVF*国际计算机视觉大会论文集，3897–3907。

Yang, Z.; Li, B.; Chen, P.-Y.; and Song, D. 2018年. 使用时间依赖性刻画音频对抗样本. *arXiv*预印本*arXiv:1809.10875*。

Yao, Z.; Wu, D.; Wang, X.; Zhang, B.; Yu, F.; Yang, C.; Peng, Z.; Chen, X.; Xie, L.; and Lei, X. 2021年. Wenet：面向生产的流式和非流式端到端语音识别工具包. *arXiv* 预印本 *arXiv:2102.01547*。

Yuan, H.; Chen, Z.; Lin, Z.; Peng, J.; Fang, Z.; Zhong, Y.; Song, Z.; 和 Gao, Y. 2025a. SatSense：用于频谱感知的多卫星协作框架。*IEEE*认知通信与网络汇刊

Yuan, H.; Chen, Z.; Lin, Z.; Peng, J.; Zhong, Y.; Hu, X.; Xue, S.; Li, W.; 和 Gao, Y. 2025b. 在低地球轨道卫星网络中利用有限样本构建四维无线电地图。*IEEE INFOCOM*。

Yuan, X.; Chen, Y.; Zhao, Y.; Long, Y.; Liu, X.; Chen, K.; Zhang, S.; Huang, H.; Wang, X.; and Gunter, C. A. 2018年{CommanderSong}：一种实用对抗语音识别的系统化方法。载于第*27*届 *USENIX* 安全研讨会（*USENIX Security 18*），49–64。

Zhang, J.; Wang, J.; Wang, H.; and Luo, X. 2022。自恢复对抗样本：社交网络中一种新的有效保护机制。*IEEE Transactions onCircuitsand SystemsforVideo Technology*，33(2)：562–574。Zhang, Y.; Chen, H.; Lin, Z.; Chen, Z.; and Zhao, J. 2025a。LCFed：一种面向异构数据的高效聚类联邦学习框架。*arXiv*预印本 *arXiv:2501.01850*。Zhang, Y.; Lin, Z.; Chen, Z.; Fang, Z.; Zhu, W.; Chen, X.; Zhao, J.; and Gao, Y. 2024。Satfed：一种资源高效的低地球轨道卫星辅助异构联邦学习框架。*Engineering*。Zhang, Z.; Duan, T.; Lin, Z.; Huang, D.; Fang, Z.; Sun, Z.; Xiong, L.; Liang, H.; Cui, H.; and Cui, Y. 2025b。面向鲁棒深度强化学习的状态感知扰动优化。*IEEE Trans.MobileComput.*Zhang, Z.; Duan, T.; Lin, Z.; Huang, D.; Fang, Z.; Sun, Z.; Xiong, L.; Liang, H.; Cui, H.; Cui, Y.; et al. 2025c。通过自适应梯度掩码对抗攻击实现机器人学中的鲁棒深度强化学习。*arXiv*预印本 *arXiv:2503.20844*。Zhao, Z.; Dua, D.; and Singh, S. 2017年。生成自然对抗样本。*arXiv*预印本 *arXiv:1710.11342*。Zhao, Z.; Liu, Z.; and Larson, M. 2021。论成功与简洁：重新审视可迁移定向攻击。神经信息处理系统进展，34：6115–6128。Zhu, J.; Du, X.; Zhou, J.; Pun, C.-M.; Xu, Q.; and Liu, X. 2024。Dp-rae：一种用于图像隐私保护的双阶段融合可逆对抗样本。载于第*32*届 *ACM*国际多媒体会议论文集，671–680。