软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(7):3033−3048 [doi: 10.13328/j.cnki.jos.007574] [CSTR: 32375.14.jos.007574]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

说话人信息引导的高性能音频对抗攻击*

陈家源, 黄文弘, 黄方军

(中山大学 网络空间安全学院, 广东 深圳 518107)

通信作者: 黄方军, E-mail: huangfj@mail.sysu.edu.cn

摘　要: 随着音频对抗攻击研究的深入, 如何在确保对抗音频隐蔽性 (即与原始音频在听觉上高度相似) 的同时, 提高其在不同模型之间的迁移性, 已成为研究热点之一. 提出一种能够同时提高对抗音频隐蔽性和迁移性的方法 SIAttack (speak information attack). 该方法的核心思想是解耦音频中的说话人信息与内容信息, 并仅对说话人信息施加轻微扰动, 从而可以在保持内容信息不变的前提下实现对说话人识别系统的高效攻击. 在 4 个说话人识别模型以及 3 个主流商业 API 上的实验表明, SIAttack 生成的音频在听觉上几乎无法与原始音频区分, 且能以较高的成功率误导所有测试模型, 在说话人识别模型上迁移成功率最高可达 100%.

关键词: 音频; 隐蔽性; 迁移性; 对抗攻击; 对抗扰动

中图法分类号: TP309

中文引用格式: 陈家源, 黄文弘, 黄方军. 说话人信息引导的高性能音频对抗攻击. 软件学报, 2026, 37(7): 3033–3048. http://www.jos.org.cn/1000-9825/7574.htm

英文引用格式: Chen JY, Huang WH, Huang FJ. High-performance Audio Adversarial Attacks Guided by Speaker Information. Ruan Jian Xue Bao/Journal of Software, 2026, 37(7): 3033–3048 (in Chinese). http://www.jos.org.cn/1000-9825/7574.htm

High-performance Audio Adversarial Attacks Guided by Speaker Information

CHEN Jia-Yuan, HUANG Wen-Hong, HUANG Fang-Jun

(School of Cyber Science and Technology, Sun Yat-sen University, Shenzhen 518107, China)

Abstract: As the research on audio adversarial attacks advances, improving the transferability of adversarial audio across different models and ensuring its imperceptibility (that is, highly similar to the original audio in auditory perception) at the same time have become a research hotspot. This study proposes a new method called speak information attack (SIAttack) that can simultaneously improve the imperceptibility and transferability of adversarial audio. Specifically, the core idea of this method is to decouple speaker information from content information in the audio, and then apply small perturbations only to the speaker information, thereby achieving efficient attacks on the speaker recognition system under the premise of keeping the content information unchanged. The experiments on four speaker recognition models and three mainstream commercial APIs show that the audio generated by SIAttack is almost indistinguishable from the original audio, and can mislead all test models with a high success rate. Additionally, the transfer success rate on speaker recognition models can reach up to 100%.

Key words: audio; imperceptibility; transferability; adversarial attack; adversarial perturbation

说话人识别系统通过提取和比对个人独特的声纹特征, 实现对个人身份的识别或验证. 近年来, 随着人工智能技术的飞速发展, 基于 AI 的说话人识别系统已广泛应用于智能家居[1]、支付交易[2]以及伪造检测[3]等现实场景, 成为现代身份认证体系中的重要组成部分. 然而, 随着该类系统的普及, 所面临的潜在安全风险也日益突出, 尤其是对抗攻击对系统可靠性构成的严峻挑战[4,5]. 攻击者可通过向音频信号中注入人耳难以察觉的对抗扰动, 误导说话人识别系统产生错误判断, 从而严重威胁系统安全[6], 引发隐私泄露、金融诈骗和法律纠纷等严重后果.

近年来, 研究者陆续提出了一系列针对说话人识别系统的对抗攻击方法. 这类方法通常通过对原始音频注入

* 基金项目: 国家自然科学基金 (U2336208); 深圳市科技计划 (JCYJ20250604175534044)
收稿时间: 2024-06-14; 修改时间: 2025-04-29; 采用时间: 2025-11-20; jos 在线出版时间: 2026-02-11
CNKI 网络首发时间: 2026-02-12

人耳难以察觉的特殊噪声, 生成对抗音频样本 (简称为对抗音频), 从而可误导识别系统, 实现身份欺诈. Kreuk 等人[7]首次对说话人验证系统实施了对抗攻击, 标志着音频对抗攻击这一研究方向的兴起. 此后, 针对端到端说话人识别系统的对抗攻击逐渐受到关注. Li 等人[8]在真实场景下实现了对说话人识别系统的攻击, 证实了音频对抗攻击在现实环境中的可行性. Xie 等人[9]则提出了一种实时、通用且鲁棒性强的对抗攻击方法. 此外, 为应对黑盒设置下的攻击难题, Chen 等人[10]提出了基于自然进化策略的黑盒对抗攻击方法.

随着音频对抗攻击研究的不断深入, 提升对抗音频的隐蔽性与迁移性, 已成为攻击方法研究中关注的重点. 隐蔽性要求对抗音频在听觉上与原音频尽可能接近, 难以被人类感知所区分; 而迁移性则强调生成的对抗音频能够同时误导多个不同的说话人识别模型. 目前, 提升对抗音频隐蔽性的方法主要可分为 3 类: 基于心理声学模型的方法[11,12]、基于听觉范围的方法[13,14]和基于特定背景噪声的方法[15,16]. 基于心理声学模型的方法利用人耳的听觉掩蔽原理[17], 将扰动隐藏在不易被感知的频率或强度范围内; 基于听觉范围的方法[13,14]将扰动添加至人耳听觉范围之外的频段 (如超声波), 从而避免被察觉; 基于特定背景噪声的方法[15,16]则将扰动伪装成常见的环境声音, 如消息提示音、键盘敲击声等, 使其在听觉上更自然. 在迁移性提升方面, 主流方法可分为两类: 一类是基于数据增强的方法[18], 另一类是基于模型集成的方法[19,20]. 基于数据增强的方法[18]在生成对抗音频时引入随机性变换, 如音频截断、填充、语速与音高调整以及添加噪声等, 以增强其在不同模型间的泛化能力; 基于模型集成的方法[19,20]在对抗音频生成过程中同时优化多个模型, 融合其梯度或决策信息, 从而提高生成样本对未知模型的攻击成功率.

尽管上述方法在提升对抗音频的隐蔽性与迁移性方面取得了一定成效, 但仍存在若干局限性: 1) 现有提高隐蔽性的策略在实际应用中面临明显限制, 如基于心理声学模型的方法未能充分考虑个体听觉感知能力的差异, 基于特定背景噪声的方法需依赖音乐、键盘声等特定声学载体等[21]; 2) 迁移性增强方法往往以牺牲隐蔽性为代价, 如为提高对抗音频在不同模型间的泛化能力, 现有方法通常需要引入更大的扰动容量, 这在一定程度上增加了对抗音频被感知或检测的风险; 3) 现有方法普遍局限于原始音频波形层面的扰动添加, 未充分探索在音频深层语义特征 (如说话人身份信息) 上进行扰动的可能.

为此, 本文提出了一种说话人信息攻击 (speaker information attack, SIAttack) 的音频对抗攻击方法. 该方法基于对音频中说话人信息与内容信息的解耦, 有针对性地对说话人信息进行扰动操作, 从而在保持高隐蔽性的同时, 提升对抗音频的迁移能力. 与传统在原始音频波形上添加扰动的方法不同, SIAttack 作用于更深层的说话人身份特征. 原始波形是低维表征, 直接扰动易产生易于检测的高频噪声; 而说话人信息是高维语义特征, 承载了声纹、口音等独特标识. SIAttack 在该高维空间引入语义一致性扰动, 能更好地保持音频自然度, 隐蔽性更强. 此外, 由于说话人身份特征是识别模型决策的核心依据, 且在不同模型间具有通用性, 攻击此特征能从根本上误导模型, 使生成的对抗音频具备强大的跨模型迁移能力, 对现有说话人识别系统构成了更深层次的安全威胁.

SIAttack 的具体实现流程如图 1 所示. 对于一段来自说话人 Bob 的原始音频, 首先将其分解为内容信息与说话人信息两部分. 内容信息主要承载语音的文本转录语义, 而说话人信息则包含用于区分说话人身份的声纹特征. 随后, 在说话人信息上施加一个微小的对抗扰动, 再结合原始内容信息与扰动后的说话人信息, 重构生成新的音频. 该音频在听觉上与原音频高度相似, 却能够有效地误导说话人识别系统做出错误判断. 实验结果表明, SIAttack 生成的对抗音频在迁移性方面表现优异, 最高可实现 100% 的攻击成功率; 同时, 人类听觉测试也证实了其良好的隐蔽性, 绝大多数测试者均未能察觉音频中存在的扰动.

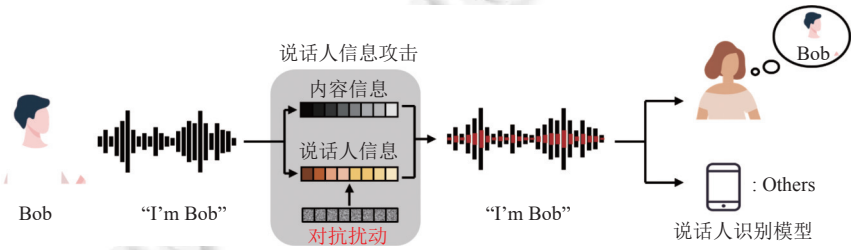


图 1 SIAttack 实现流程图

本文第 1 节简要介绍说话人识别系统. 第 2 节详细介绍本文方法 SIAttack. 第 3 节介绍实验分析. 第 4 节介绍

相关讨论. 第 5 节总结本文.

1 说话人识别系统

当前, 说话人识别系统主要包含注册与识别两个阶段. 在注册阶段, 设有一个由 n 个说话人构成的说话人集合为 $G=\{1,2,\ldots,n\}$. 系统首先利用背景模型将这些说话人的音频映射为对应的注册嵌入 (通常表示为一维向量). 背景模型可采用高斯混合模型[22]、i-vector[23]等传统模型, 也可选用如 d-vector[24]、x-vector[25]等深度神经网络模型. 在识别阶段, 对于一段未知音频 x, 系统同样通过背景模型提取其嵌入表示, 随后利用评分模块 $S(\cdot)$ 计算该嵌入与 n 个说话人的注册嵌入之间的匹配分数 $S(x)\in\mathbb{R}^n$, 最终依据得分情况输出识别结果.

说话人识别系统通常可完成两类任务: 闭集识别任务 (close-set identification, CSI) 和开集识别任务 (open-set identification, OSI). 在 CSI 任务中, 系统仅需在已知的说话人集合 G 中选择最可能的说话人, 而无需考虑集合外的其他说话人. 具体而言, 对于输入音频 x, 系统首先计算其与 G 中所有说话人的匹配分数 $S(x)$, 随后直接选择分数最高的说话人作为识别结果 $J(x)$, 即 $J(x)_{\mathrm{CSI}}=\underset{i\in G}{MaxIndex}(S(x)_i)$, 其中 $MaxIndex(\cdot)$ 返回向量中最大值所对应的索引. 而在 OSI 任务中, 系统不仅需判断 x 属于 G 中的哪一个说话人, 还需考虑其是否可能为未知的冒充者. 具体流程中, 系统在计算匹配分数后, 会进一步结合预设的分数阈值 θ 来判断音频 x 进行如下决策, $J(x)_{\mathrm{OSI}}=-1$ 时为冒充者.

$$J(x)_{\mathrm{OSI}}=\begin{cases}\underset{i\in G}{MaxIndex}(S(x)_i), & \text{if } \max\limits_{i\in G}S(x)_i\geqslant\theta\\ -1, & \text{otherwise}\end{cases}\tag{1}$$

2 基于说话人信息的对抗攻击

2.1 场景设定

本文构建了白盒攻击与黑盒攻击两类实验场景, 以全面评估不同对抗攻击算法的性能. 在白盒攻击场景中, 攻击者完全掌握目标模型的内部信息, 包括模型架构、参数、训练算法及所用数据集等. 在黑盒攻击场景下, 攻击者无法获知目标模型的任何内部信息, 仅能通过模型输入输出接口访问并获取预测结果. 在黑盒条件下, 常见的攻击手段主要包括基于迁移的方法和基于查询的方法两类: 基于迁移的方法通过在替代模型上生成对抗音频, 并依赖其跨模型迁移能力对黑盒目标实施攻击; 基于查询的方法则通过多次调用目标模型接口, 根据反馈逐步优化对抗音频. 然而, 实际商业环境中, 多数说话人识别服务 (如京东语音 API) 会对用户查询频率和次数施加严格限制, 例如每秒最多 2 次查询, 每日上限 500 次, 并可能收取相应查询费用, 这使得基于查询的攻击成本显著增加. 为贴近真实攻击条件, 本文设定所有基于查询的黑盒攻击算法在每个对抗音频的生成过程中最多可进行 500 次查询.

不失一般性, 假设攻击者能够利用当前丰富的开源语音数据集训练替代模型. 同时, 攻击者已知目标系统中注册说话人的身份信息, 并可通过公开渠道 (如社交媒体或现场录音) 获取该说话人的部分音频样本, 以供后续攻击使用. 已有研究[26]表明, 源说话人与目标说话人的不同组合以及识别任务的差异可构成多种攻击场景. 根据文献 [19] 的建议, 本文重点研究 OSI 任务下的两种典型攻击情景: 无目标攻击 (OSI-u) 与有目标攻击 (OSI-t). 在这两种情景中, 攻击者均试图利用冒充者的原始音频生成对抗音频, 以误导目标说话人识别系统产生错误判断. 二者的区别在于攻击目标不同: OSI-u 旨在使对抗音频被系统错误地接受为任意一个已注册说话人; 而 OSI-t 则力求使样本被识别为某个特定的目标注册说话人. 考虑到 OSI 任务在实际应用中比 CSI 任务更具挑战性, 且目前尚无主流商业说话人识别系统采用 CSI 任务架构[19], 本文将研究重点集中于更具现实意义的 OSI 任务场景.

2.2 对抗音频生成

传统对抗攻击方法核心为构造一个人耳难以察觉的微小扰动 δ, 将其添加至干净音频 x, 以得到对抗音频 x', 即:

$$x'=x+\delta\tag{2}$$

然而, 这种直接在原始音频上添加扰动的方法, 往往难以兼顾对抗样本的隐蔽性与迁移性. 为此, 本文提出了 SIAttack 框架 (如图 2 所示), 其核心思想是通过在音频 x 的说话人信息上添加一个微小的扰动 δ, 以生成具有高隐

© 中国科学院软件研究所 http://www.jos.org.cn

蔽性和迁移性的对抗音频 x'. 该过程主要分为两步: 1) 原音频信息解耦; 2) 修改说话人信息并重建音频.

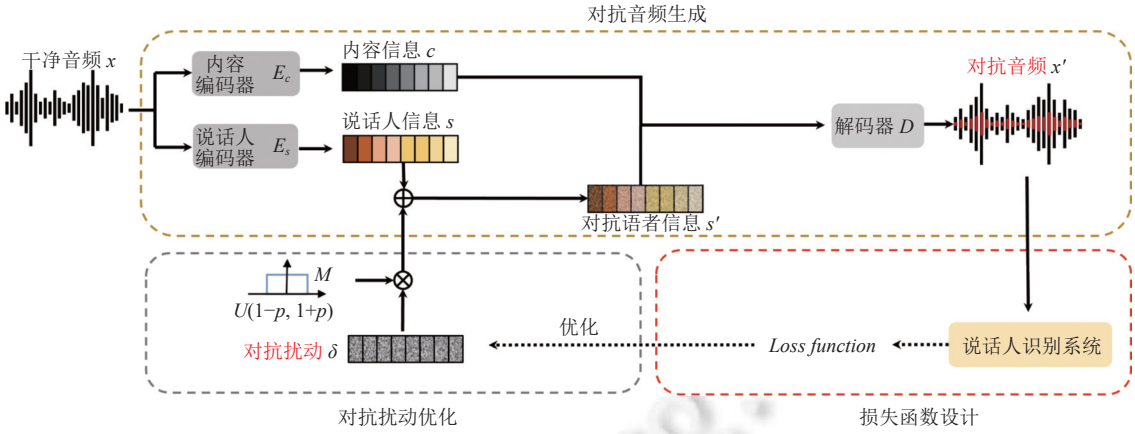


图 2 SIAttack 的整体框架

1) 原音频信息解耦: 如图 2 所示, 将干净音频 x 同时输入一个内容编码器 (E_c) 和一个说话人编码器 (E_s), 分别得到内容信息 c 和说话人信息 s. 其中, 内容编码器[27]由 WavLM 模型和瓶颈提取器组成. WavLM 模型以原始波形为输入, 首先生成包含内容信息和说话人信息的高维表示. 随后, 将这些高维表示放入瓶颈提取器中转换为低维表示. 这种巨大的维度差距可以造成信息瓶颈, 从而迫使产生的低维表示放弃与内容无关的信息, 如噪声或说话人信息. 说话人编码器则是一个在大量的语音数据集上预训练的说话人验证模型, 能有效提取说话人的身份信息. 本文采用的说话人编码器[28]采用了基于长短期记忆网络 (long short-term memory, LSTM) 的模型架构, 总共包含 3 层, 每层有 256 个隐藏节点. 其后接有一个包含 256 个单元的投影层, 最终对最后一层的隐藏状态进行 L2 规范化, 所得向量即为说话人嵌入表示.

2) 修改说话人信息并重建音频: 在提取的说话人信息 s 后, 为其添加一个微小对抗扰动 δ. 为增强生成对抗音频的迁移性, 将 δ 与一个随机矩阵 $\mathbf{M}$ 进行 Hadamard 运算处理. 矩阵 $\mathbf{M}$ 中的每个元素独立地从均匀分布 $U(1-p, 1+p)$ 中采样生成, 其作用是在扰动 δ 中引入随机性和复杂性, 以有效减轻对抗音频在源模型上的过拟合, 从而提升其在不同目标模型间的迁移能力. 参数 p 用以控制均匀分布中采样值的范围. 当 p 值越大, 采样值的波动范围越广, 引入的随机性也越强. 如公式 (2) 所示, 扰动处理后的说话人信息 s' 可表示为:

$$s' = s + \delta \odot \mathbf{M} \tag{3}$$

接下来, 将内容信息 c 与扰动后的说话人信息 s' 共同输入解码器 $D(\cdot,\cdot)$ 中, 重构生成对抗音频 $x' = D(c, s')$. 本文所使用的解码器 $D(\cdot,\cdot)$ 结构与文献 [29] 中相同, 由多个反向卷积层组成, 每个卷积层后接有一个多接受场融合模块.

2.3 损失函数设计

为实现对抗音频在有效误导说话人识别系统的同时保持良好隐蔽性, 设计了如下损失函数用于优化扰动 δ:

$$\mathcal{L}_{\mathrm{SIA}} = \lambda_1 \cdot \mathcal{L}_{\mathrm{adv}} + \lambda_2 \cdot \mathcal{L}_{\mathrm{per}} \tag{4}$$

其中, $\mathcal{L}_{\mathrm{adv}}$ 代表对抗损失, 旨在增强对抗音频的攻击性, 误导系统产生错误分类; $\mathcal{L}_{\mathrm{per}}$ 为感知损失, 用于确保对抗音频与原始音频在感知上高度相似. 两者通过权重系数 λ_1 和 λ_2 进行平衡.

对抗音频 x' 的目标是使得说话人识别系统将其识别为任意其他注册说话人发出的, 即 $J(x') \neq J(x)$; 或者将其识别为特定目标注册说话人发出的, 即 $J(x') = t$, t 为目标说话人标签. 为了实现上述无目标攻击和目标攻击, 需使对抗音频 x' 与说话人组 G 中的说话人的匹配分数尽可能提高. 因此, $\mathcal{L}_{\mathrm{adv}}$ 定义如下:

$$\mathcal{L}_{\mathrm{adv}} = \begin{cases} \theta - \max_{i \in G} S(x')_i + \kappa, & \text{Untargeted} \\ \max\{\theta, \max_{i \in G, i \neq t} S(x')_i\} - S(x')_t + \kappa, & \text{Targeted} \end{cases} \tag{5}$$

其中, $\max_{i \in G, i \neq t} S(x')_i$ 代表对抗音频 x' 与注册说话人组 G 内, 除目标说话人外其他说话人匹配分数的最大值, κ 为置

© 中国科学院软件研究所 http://www.jos.org.cn

信度, θ 为前文所介绍的 OSI 任务中的分数阈值.

为确保生成扰动后的说话人信息 s' 与原说话人信息 s 尽可能相似, 使用余弦相似度计算感知损失 $\mathcal{L}_{\text{per}}$:

$$\mathcal{L}_{\text{per}} = 1 - Sim(s, s') = 1 - \frac{s \cdot s'}{\|s\| \|s'\|} \tag{6}$$

其中, $Sim(s, s')$ 用于计算 s 和 s' 之间的余弦相似度.

2.4 对抗扰动优化

根据以上构造的损失函数 $\mathcal{L}_{\text{SIA}}$, 采用迭代梯度下降的方式优化 δ, 即:

$$\delta_{i+1} \leftarrow clip_{\varepsilon}(\delta_i - lr \cdot sign(\nabla_{\delta} \mathcal{L}_{\text{SIA}})) \tag{7}$$

其中, lr 是学习率, $sign(\cdot)$ 是符号函数, $clip_{\varepsilon}(\cdot)$ 函数对扰动幅值进行限定, 即 $\|\delta\|_{\infty} < \varepsilon$. 优化步数为 N, 为了缩短对抗音频生成时间, 实验中使用了文献 [10] 中的早停策略, 即对抗音频成功误导目标模型时便立刻停止优化. SIAttack 的整体过程可用算法 1 表示.

算法 1. SIAttack.

输入: 干净音频 x;
输出: 对抗音频 x'.

1. 提取 x 的内容信息 c
2. 提取 x 的说话人信息 s
3. 计算干净音频 x 的匹配分数 $J(x)$
4. 初始化扰动 δ_0 为 0
5. **for** $i = 0$ **to** N **do**
6. 　构造添加了扰动的说话人信息 $s' = s + \delta_i \odot \mathbf{M}$
7. 　重建新音频 $x' = D(c, s')$
8. 　计算对抗音频 x' 的匹配分数 $J(x')$
9. 　**if** $J(x') \neq J(x)$ (有目标攻击则为 $J(x') = t$) **then**
10. 　　**return** x'
11. 　**else**
12. 　　计算损失 $\mathcal{L}_{\text{SIA}}$
13. 　　更新扰动 $\delta_{i+1} \leftarrow clip_{\varepsilon}(\delta_i - lr \cdot sign(\nabla_{\delta} \mathcal{L}_{\text{SIA}}))$

2.5 扰动阈值评估

在生成对抗音频的过程中, 解码器 $D(\cdot,\cdot)$ 根据扰动后的说话人信息重建音频. 若添加的扰动过大, 可能导致重构音频在听觉上表现为另一说话人的音色. 因此, 在扰动更新时 (公式 (7)), 需确保 $\|\delta\|_{\infty} < \varepsilon$, 从而使得重构的对抗音频 x' 与原音频 x 尽可能相似. 然而, 仅凭经验设定阈值 ε 难以取得理想效果. 为此, 本文提出一种扰动阈值自动评估算法, 以更合理地确定该阈值. 该算法的核心思想是: 首先将扰动阈值 ε 初始化为一个较大值, 随后逐步衰减, 直至重构音频与原始音频在说话人音色上的相似度达到可接受水平时, 迭代终止. 通过计算多个音频的平均阈值 $\bar{\varepsilon}$ 作为最终设定的 ε 值, 每个音频的阈值评估步骤如下.

1) 对于音频 x_i ($i = \{1, \ldots, n\}$), 将其输入内容编码器和说话人编码器, 得到内容信息 c_i 和说话人信息 s_i. 扰动阈值 ε_i 初始设置为 $\max(s_i) - \min(s_i)$.

2) 构造扰动 δ_i, 其值从均匀分布 $U(-\varepsilon_i, \varepsilon_i)$ 中采样得到, 添加至 s_i 后得到 $s'_i = s_i + \delta_i$, 通过解码器重建出新音频 x'_i.

3) 计算重建音频 x'_i 与原始音频 x_i 的频谱相似度 Cs.

4) 判断 Cs 是否大于 0.7, 如果否, 则令 $\varepsilon_i = \varepsilon_i \cdot 0.8$, 以降低其大小并跳转至第 2) 步.

© 中国科学院软件研究所 http://www.jos.org.cn

5) 输出此时扰动阈值 ε_i.

3 实验分析

本节对 SIAttack 及多种代表性的对抗攻击方法进行全面评估, 设计了 5 项实验: 1) 白盒攻击, 在完全了解目标模型时评估攻击成功率; 2) 黑盒攻击, 基于迁移性评估对未知模型的攻击效果; 3) 商业 API 攻击, 在真实商业系统中验证方法的实际威胁; 4) 人类听觉评估, 通过主观实验检验对抗音频的隐蔽性与自然度; 5) 防御鲁棒性, 评估所生成样本对抗现有主流防御手段的能力.

3.1 实验设置

本文采用 LibriSpeech[30]和 VCTK[31]作为数据来源. 鉴于原始数据规模较大, 我们从中筛选部分音频样本, 构建了两个精简子集 LS (源自 LibriSpeech) 与 VT (源自 VCTK), 具体划分如表 1 所示. LS 数据集进一步细分为 LS_{enroll} 和 LS_{OSI} 两个子集. 类似地, VT 数据集也划分为 VT_{enroll} 和 VT_{OSI} 两个子集. 其中, 标注为“enroll”的子集用于说话人识别系统的注册阶段; 标注为“OSI”的子集用于在开集识别任务中生成与测试对抗音频.

表 1 实验数据集划分

数据集	音频数量(个)	说话人数量	子数据集	用途	音频数量 (个)	数据来源
LS	1 100	10	LS_{enroll}	注册	100	LibriSpeech
			LS_{OSI}	OSI任务	1 000	
VT			VT_{enroll}	注册	100	VCTK
			VT_{OSI}	OSI任务	1 000	

注: LS_{enroll}和LS_{OSI}这2个子集均从文献[32]中获得

实验共采用以下 4 个开源说话人识别系统作为目标模型: IV_PLDA[33]、XV_PLDA[34]、ECAPA[35]和 ResNet[36]. 模型详细信息如表 2 所示. 这些模型在架构设计、输入声学特征类型及识别阶段所使用的评分模块等方面均存在显著差异, 为后续攻击效果的对比分析提供了充分的多样性基础. 所有参与实验的模型都被验证能够有效完成 LB 数据集和 VT 数据集上的说话人识别任务. 为进一步评估攻击方法在真实场景下的适用性, 本文还引入了 3 家主流商业声纹识别 API 作为测试对象, 分别为腾讯语音识别 (Tencent)[37]、科大讯飞声纹识别 (iFlytek)[38]和云知声声纹识别 (Unisound)[39]. 所有模型 (包括开源与商业系统) 在正常环境下的识别准确率及阈值 θ 设定如表 3 所示.

表 2 目标模型的详细信息

模型架构	模型名称	参数量 (M)	声学信息	训练数据集	评分模块
GMM	IV_PLDA	80.37	MFCC	LibriSpeech	PLDA
TDNN	ECAPA	20.77	fBank	VoxCeleb1&2	COSS
	XV_PLDA	5.79	MFCC	LibriSpeech	PLDA
CNN	ResNet	11.17	Spectrogram	VoxCeleb1&2	COSS

表 3 目标系统的说话人识别准确率

数据集	IV_PLDA		XV_PLDA		ECAPA		ResNet		Tecent		iFlytek		Unisound	
	θ	准确率 (%)	θ	准确率 (%)	θ	准确率 (%)	θ	准确率 (%)	θ	准确率 (%)	θ	准确率 (%)	θ	准确率 (%)
LS_{OSI}	16.3	97.6	16.4	97.6	0.42	99.6	0.45	96.6	0.7	96.4	0.6	97.3	0.6	98.3
VT_{OSI}	16.3	93.5	16.4	92.5	0.42	97.9	0.45	98.9	0.7	96.3	0.6	96.5	0.6	98.4

实验中使用攻击成功率 (即说话人识别系统被对抗音频成功误导的比率) 来评估各种攻击方法的性能, 使用信噪比 (signal-to-noise ratio, SNR) 和音频质量感知评估 (perceptual evaluation of speech quality, PESQ)[40] 指标分别量化扰动幅度与听觉质量. SNR 值越高, 说明扰动越不易被察觉. PESQ 模拟人类听觉系统的客观感知测量, PESQ 分值越高, 表示听觉质量越接近原始音频.

© 中国科学院软件研究所 http://www.jos.org.cn

本文对比了多种攻击方法, 包括 4 种从图像领域迁移至音频领域的经典对抗攻击方法, 即 FGSM[41]、PGD[42]、CW2[43]和 CWinf[43], 以及 6 种音频对抗攻击方法, 分别为 Inaudible[12]、SirenAttack[44]、FakeBob[10]、Kenansville[45]、QFA2SR[19]和 SMACK[46]. 其中, SirenAttack、FakeBob 和 SMACK 均属于基于查询的黑盒算法. 本文所提出的 SIAttack 在实验中的参数设置如下: 均匀分布系数 $p = 0.05$、最大迭代次数 $N = 1000$、学习率 $lr = 0.001$、两个损失函数系数 $\lambda_1 = 1$ 和 $\lambda_2 = 0.1$、扰动阈值 $\varepsilon = 0.084$. 置信度 κ 的设定与目标模型的阈值有关, 当目标模型为 IV_PLDA 和 XV_PLDA 时, $\kappa = 3$; 其余情况下, $\kappa = 0.1$. CW2、FakeBob 和 SirenAttack 的置信度设置与此相同. 另外, 为了验证不同扰动载体对攻击性能的影响, 我们还同时生成了在内容信息上添加扰动的对抗音频. 后文使用 SIAttack-c 和 SIAttack-s 分别表示在内容信息和说话人信息添加扰动的对抗音频.

3.2 白盒场景下的攻击性能

在白盒测试环境中, 所有攻击方法均可利用目标模型提供的完整梯度信息进行扰动优化. 表 4 展示了各方法在 LS 数据集上的攻击性能结果.

表 4 LS 数据集下基于迁移方法在白盒场景的攻击性能

Attack	成功率 (%)								SNR (dB)	PESQ	Time (s)
	IV_PLDA		XV_PLDA		ECAPA		ResNet				
	OSI-u	OSI-t	OSI-u	OSI-t	OSI-u	OSI-t	OSI-u	OSI-t			
FGSM	75.2	30.1	84.6	28.5	86.5	24.0	71.5	26.0	32.5	2.5	**0.04**
PGD	**100.0**	98.5	**100.0**	**100.0**	**100.0**	99.4	99.7	99.6	36.1	3.1	3.0
CW2	**100.0**	96.3	**100.0**	**100.0**	95.3	86.0	82.5	75.0	**51.3**	**4.2**	79.6
CWinf	**100.0**	100.0	**100.0**	**100.0**	**100.0**	87.0	99.8	85.0	34.2	2.8	3.2
Inaudible	**100.0**	91.0	98.2	92.1	95.0	65.5	96.6	85.2	36.1	4.1	85.5
Kenansville	75.6	—	72.6	—	56.2	—	64.5	—	10.1	1.4	2.0
QFA2SR	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	25.7	2.8	12.2
SIAttack-c	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	25.5	4.1	10.4
SIAttack-s	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	24.3	4.1	10.2

注: 加粗数据表示最优结果

根据表 4 所示的攻击成功率结果, 可以观察到不同攻击方法在性能上呈现出显著差异. 由于模型架构的不同, 针对 IV_PLDA 和 XV_PLDA 模型的攻击成功率高, 而对 ECAPA 和 ResNet 模型的攻击则较为困难. 此外, 受任务难度影响, 大多数方法在 OSI-u (无目标攻击) 任务中表现较好, 而在更具挑战性的 OSI-t (有目标攻击) 任务中成功率普遍偏低. 需要说明的是, Kenansville 方法仅适用于无目标攻击. 值得关注的是, 在多数实验设置下, QFA2SR、SIAttack-c 和 SIAttack-s 这 3 种方法均实现了 100% 的攻击成功率, 这主要得益于其采用了更强的攻击策略并引入了更大的扰动幅度. 在扰动设置方面, FGSM、PGD 和 CWinf 等常规攻击方法将扰动阈值设定为 0.02, QFA2SR 未设置类似阈值, SIAttack 的扰动阈值则为 0.086. 通过 SNR 指标可以进一步量化不同方法引入的扰动大小: CW2 具有最高的 SNR 值, 表明其产生的扰动最小; 而 QFA2SR 和 SIAttack 的 SNR 值较低, 说明它们引入了相对较大的扰动.

在音频质量方面, QFA2SR 由于引入了较大的扰动, 其 PESQ 得分较低, 表明所生成的对抗音频的听觉质量受到明显影响. 相比之下, SIAttack 取得了更高的 PESQ 评分, 反映出其生成的对抗音频具有更优的听觉保真度. 这一优势主要源于 SIAttack 在内容信息和说话人信息等高维语义空间中施加扰动, 其扰动本身带有语义一致性; 而传统方法多直接在原始波形上添加类似于高频噪声的扰动, 这种非语义的干扰方式更容易损害音频的听觉自然度. 需要指出的是, 尽管 CW2 在 SNR 和 PESQ 两项客观指标上均表现最佳, 但这并不完全等同于其在人耳听感中具备最优的音频质量, 相关主观听觉测试结果将在第 3.5 节中进一步讨论.

在计算效率方面, FGSM 表现最佳, 每个对抗音频仅需一次迭代即可生成, 平均耗时仅为 0.04 s. 相比之下, CW2 与 Inaudible 等方法由于采用二分搜索等优化策略, 所需迭代次数较多, 生成时间显著延长. PGD 方法需进行 100 次迭代, 平均耗时为 3.0 s; 而 SIAttack 平均需要 300 次迭代 (详见第 4.3 节), 平均生成时间达到 10.3 s. 这一方

© 中国科学院软件研究所 http://www.jos.org.cn

面源于迭代次数的增加, 另一方面也受到音频解耦与重建过程中额外计算开销的影响. 通用的算法-硬件联合优化与部分专用的加速架构[47]在理论上可以缩短单次迭代的时间, 但依然需要多次迭代以生成对抗音频. 值得注意的是, SIAttack-c 的生成时间略长于 SIAttack-s, 这主要是因为内容信息的数据维度高于说话人信息. 通常情况下, 内容信息是一个二维向量 (时间长度与特征维度), 例如本研究中一段 10 s 音频的内容信息维度为 1024×690; 而说话人信息为固定长度的一维向量 (本研究为 256 维). 因此, 在说话人信息层面添加扰动能够有效减少计算负载, 从而提高对抗音频的生成效率.

为研究数据集对攻击性能的影响, 我们进一步在 VT 数据集上评估了各方法的攻击性能 (结果见表 5). 实验结果显示, 尽管各方法在 VT 数据集上的表现趋势与之前基本一致, 但整体攻击成功率出现了一定程度的下降. 由表 3 可知, 目标模型在 LS 数据集上的分类准确率普遍高于 VT 数据集.

表 5 VT 数据集下基于迁移方法在白盒场景的攻击性能

Attack	成功率 (%)								SNR (dB)	PESQ	Time (s)
	IV_PLDA		XV_PLDA		ECAPA		ResNet				
	OSI-u	OSI-t	OSI-u	OSI-t	OSI-u	OSI-t	OSI-u	OSI-t			
FGSM	69.6	25.0	55.0	17.0	35.0	6.1	38.0	12.5	33.0	2.8	**0.04**
PGD	**100.0**	95.6	**100.0**	92.7	96.2	89.1	97.1	91.5	37.1	3.2	3.4
CW2	**100.0**	97.2	99.8	88.5	85.9	79.3	73.6	65.4	**52.1**	**4.3**	81.1
CWinf	**100.0**	98.8	92.0	88.3	92.9	78.7	87.2	77.7	34.9	3.0	5.2
Inaudible	98.2	82.2	94.0	80.0	86.8	60.8	87.6	77.3	36.0	3.9	87.4
Kenansville	67.6	—	65.4	—	48.9	—	59.5	—	11.0	1.7	2.1
QFA2SR	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	26.2	2.9	12.6
SIAttack-c	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	24.3	4.1	10.3
SIAttack-s	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	24.5	4.2	10.1

注: 加粗数据表示最优结果

3.3 黑盒场景下的攻击性能

为更真实地模拟实际攻击条件, 本文系统评估了所有基于迁移和基于查询的攻击方法在黑盒场景下的表现. 在基于迁移的攻击中, 每次实验从 4 个开源目标模型中选取 3 个作为替代模型, 另一个作为目标模型. 采用文献 [19] 提出的 Sum-Global 模型集成策略生成对抗音频, 并将其迁移至目标模型, 进而评估其攻击成功率. 对于基于查询的攻击方法, 实验模拟了实际环境中查询次数受限的情况, 攻击者仅能通过有限次数的模型查询获取反馈以完成对抗音频的生成. 相关实验结果见表 6 与表 7.

表 6 LS 数据集下各方法在黑盒场景的攻击性能

Attack	成功率 (%)								SNR (dB)	PESQ	Time (s)
	IV_PLDA		XV_PLDA		ECAPA		ResNet				
	OSI-u	OSI-t	OSI-u	OSI-t	OSI-u	OSI-t	OSI-u	OSI-t			
FGSM	55.7	12.0	42.6	13.4	16.5	3.4	37.7	8.6	32.8	2.5	**1.5**
PGD	85.7	40.7	82.1	38.9	63.5	29.6	73.9	33.8	36.4	3.1	5.0
CW2	87.7	43.7	83.3	41.8	65.6	32.4	75.4	37.6	**51.6**	**4.2**	211.8
CWinf	85.6	45.3	81.3	42.5	64.8	34.6	72.9	37.4	34.5	2.8	10.9
Inaudible	82.5	41.1	78.7	39.4	61.2	30.7	69.9	35.3	35.3	4.0	258.1
Kenansville	76.3	—	73.1	—	56.2	—	64.7	—	10.1	1.4	8.2
QFA2SR	**100.0**	90.7	94.8	89.0	77.1	55.5	89.1	72.5	25.8	2.8	16.1
SirenAttack	80.4	65.4	77.2	62.5	60.3	48.5	68.5	54.9	30.4	2.6	189.5
FakeBob	96.0	76.2	90.9	72.6	71.5	57.7	81.6	63.6	31.0	2.8	125.0
SMACK	**100.0**	82.5	91.9	78.1	72.7	61.6	83.3	70.5	—	—	258.1
SIAttack-c	**100.0**	86.0	88.6	82.1	78.9	58.8	82.6	69.8	25.1	4.1	15.8
SIAttack-s	**100.0**	**94.5**	**97.8**	**93.8**	**90.1**	**65.4**	**92.7**	**78.9**	25.7	4.1	15.1

注: 加粗数据表示最优结果

© 中国科学院软件研究所 http://www.jos.org.cn

表 7 VT 数据集下各方法在黑盒场景的攻击性能

Attack	成功率 (%)								SNR (dB)	PESQ	Time (s)
	IV_PLDA		XV_PLDA		ECAPA		ResNet				
	OSI-u	OSI-t	OSI-u	OSI-t	OSI-u	OSI-t	OSI-u	OSI-t			
FGSM	44.5	0.0	31.0	0.0	10.5	0.8	15.9	1.5	31.4	2.4	**1.5**
PGD	82.2	35.2	72.2	33.7	55.8	25.2	64.7	29.2	35.3	3.1	5.1
CW2	80.7	37.9	74.0	36.6	58.2	28.0	66.3	32.8	**51.0**	**4.2**	207.5
CWinf	79.6	39.1	72.2	36.7	57.3	29.9	64.7	32.4	33.2	2.9	10.5
Inaudible	78.3	35.6	69.3	34.3	53.7	26.4	61.4	30.7	36.9	4.1	256.8
Kenansville	69.9	—	64.8	—	49.8	—	57.1	—	9.6	1.2	11.2
QFA2SR	94.1	88.7	93.9	87.5	76.1	54.7	87.4	70.9	25.1	2.8	16.1
SirenAttack	78.8	63.9	76.0	61.0	59.3	47.3	67.6	53.5	31.1	2.8	228.4
FakeBob	94.8	75.0	89.2	71.1	70.5	56.7	79.9	62.3	32.0	2.9	156.3
SMACK	96.3	81.5	91.2	76.9	71.3	60.5	81.9	69.3	—	—	295.1
SIAttack-c	88.4	84.3	87.2	80.0	77.2	57.5	81.3	68.3	25.5	4.1	15.7
SIAttack-s	**98.4**	**93.5**	**96.4**	**92.4**	**88.7**	**64.0**	**91.7**	**77.1**	25.9	4.1	15.2

注: 加粗数据表示最优结果

在攻击成功率方面, 与白盒场景相比, 大多数基于迁移的方法在黑盒攻击中表现有所下降. 而基于查询的黑盒攻击方法则普遍取得更高的成功率. 此外, OSI-u 与 OSI-t 任务间的难度差异更为突出, 多数方法在 OSI-t 上的成功率仅为 OSI-u 的一半左右. 不同目标模型的攻击难度也存在明显差异, 其中对 ECAPA 和 ResNet 的攻击成功率显著低于其他模型. 值得关注的是, SIAttack-c 与 SIAttack-s 在黑盒环境下仍保持了较高的攻击成功率, 显著优于其他对比方法. 这表明在音频的高维语义特征 (如内容与说话人信息) 上施加扰动, 相较于直接在原始波形上添加噪声, 能够产生迁移性更好的对抗音频. 同时, SIAttack-s 相比 SIAttack-c 在成功率上约有 10% 的提升, 说明在说话人信息层面引入扰动具有更强的通用性.

在音频质量方面, 基于查询的攻击方法所生成的对抗音频在 SNR 和 PESQ 指标上普遍偏低 (注: SMACK 方法因修改了音频的音律结构, 导致对抗音频与原始音频长度不一致, 无法有效计算 SNR 和 PESQ). 相比之下, SIAttack 的表现更优, 反映出其生成的对抗音频在保持听觉质量方面具有明显优势.

在计算效率上, 基于迁移的方法因采用模型集成策略, 生成时间有所增加; 而基于查询的方法耗时更为显著, 例如 SMACK 平均需 258 s 才能生成一个对抗音频. 这主要是由于在黑盒设置下, 攻击者无法获取模型内部信息, 只能依赖多次查询和反馈逐步调整攻击策略, 这一迭代过程自然比直接利用模型梯度的迁移攻击更为耗时. 相较之下, SIAttack 平均仅需 15 s 即可生成一个对抗音频, 展现出更高的生成效率.

3.4 针对商业模型的攻击测试

为更全面地评估各攻击方法在实际应用中的有效性, 本文进一步测试了所有方法针对 3 种主流商业说话人识别 API 的攻击效果, 包括腾讯语音识别 (Tencent) 、科大讯飞声纹识别 (iFlytek) 及云知声声纹识别 (Unisound). 与这些商业 API 的交互均通过 HTTP POST 请求实现. 考虑商业服务通常按调用次数计费, 为控制实验成本, 本研究仅使用 LS 数据集进行评估. 具体而言, 注册阶段采用 LS_{enroll} 中的所有音频样本, 测试阶段则从 LS_{OSI} 中随机抽取 100 条音频作为测试样本. 为确保基于迁移的攻击方法达到最佳性能, 实验将其替代模型扩展为全部 4 个开源模型 (IV_PLDA、XV_PLDA、ECAPA 和 ResNet), 以充分利用集成模型的迁移优势. 详细攻击结果汇总于表 8.

在攻击成功率方面, SIAttack-s 在 Tencent 和 iFlytek 两个商业 API 上表现最佳, 其在 OSI-t 任务中的攻击成功率分别比 QFA2SR 高出 8% 和 17%. 然而, 在对 Unisound API 的攻击中, QFA2SR 取得了最高成功率, 这一差异可能源于不同商业模型在结构设计与防御机制上的区别.

从计算效率角度分析, 基于查询的攻击方法在商业 API 测试中耗时显著增加, 以 SMACK 为例, 其平均生成时间从 258 s 延长至 452 s. 这主要受到两方面因素影响: 一是 HTTP 请求本身存在的网络延迟, 二是商业 API 对

© 中国科学院软件研究所 http://www.jos.org.cn

单位时间内的请求次数设有严格限制. 相比之下, SIAttack 的平均耗时仅为 17 s, 充分体现了基于迁移的攻击方法在真实场景下的效率优势.

表 8 各方法在商业模型 API 上的攻击性能

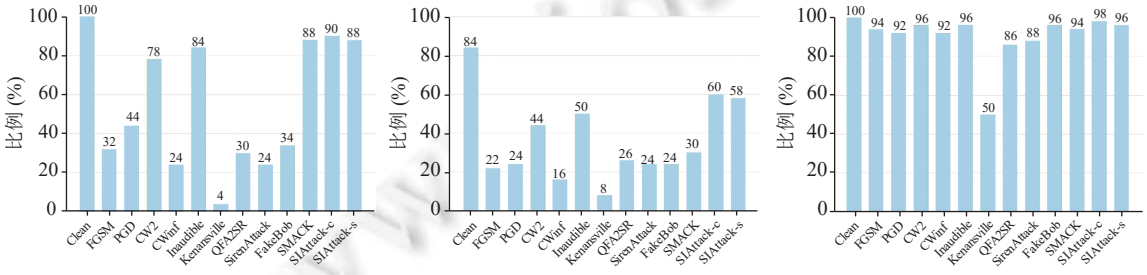
Attack	成功率 (%)						SNR (dB)	PESQ	Time (s)
	Tencent		iFlytek		Unisound				
	OSI-u	OSI-t	OSI-u	OSI-t	OSI-u	OSI-t			
FGSM	7	1	10	2	13	3	33.5	2.5	**1.7**
PGD	26	13	33	20	78	37	38.3	3.0	6.5
CW2	27	14	37	12	82	37	**50.8**	**4.2**	276.6
CWinf	26	14	35	12	80	42	34.0	2.8	12.1
Inaudible	26	13	31	19	76	37	35.2	4.1	294.0
Kenansville	21	0	37	0	74	0	9.5	1.4	13.6
QFA2SR	52	23	79	49	**93**	**85**	23.0	2.3	19.7
SirenAttack	16	4	53	15	73	26	32.7	2.3	393.8
FakeBob	36	10	58	17	79	36	29.0	2.1	224.0
SMACK	40	12	43	24	58	37	—	—	452.6
SIAttack-c	58	22	79	52	86	73	25.7	4.1	17.8
SIAttack-s	**63**	**31**	**82**	**66**	90	83	25.9	4.1	17.2

注: 加粗数据表示最优结果

3.5 人类听觉测试

隐蔽性是对抗音频的另一关键属性, 为评估各方法在此方面的表现, 本节开展了人类听觉测试. 测试样本选自第 3.4 节用于攻击商业 API 的对抗音频, 以及每种方法针对同一条原始音频所生成的对抗版本. 基于这些样本, 本文设计了 3 种对比实验, 并通过问卷调查平台[48]发放了 50 份有效问卷以收集主观听感评价.

实验 1: 判断给定音频是否不含噪声. 本实验向测试者提供一个未知音频 (任意攻击方法的生成对抗音频, 或是一个干净音频 Clean), 并让测试者判断给定的音频是否不含噪声, 结果如图 3(a) 所示, 其展示了测试者认为提供的音频不含噪声的比例. 作为参考基准, 全部参与者 (100%) 均认为 Clean 音频不含任何可察觉噪声. 其他方法生成的对抗音频在感知评价上呈现明显差异: 具体而言, 分别有 84% 和 88% 的测试者认为 Inaudible 和 SMACK 方法生成的音频不含噪声. 这归因于 Inaudible 采用了基于心理声学模型的掩蔽策略, 而 SMACK 通过对音频韵律进行结构性修改以生成对抗音频, 两者均能有效维持较好的听觉质量. 相比之下, QFA2SR 方法仅有 30% 的听感通过率, 说明其在追求高迁移性的过程中显著牺牲了音频质量. Kenansville 方法表现最不理想, 因其在信号处理过程中过度削减频段内容, 导致音频失真严重. 值得关注的是, SIAttack-c 与 SIAttack-s 分别获得 90% 和 88% 的“无噪声”判断率, 显著优于多数对比方法, 反映出本文所提方法能够在实现对抗性的同时, 更好地维持音频的自然听感.



(a) 实验 1: 测试者认为给定音频不含噪声的比例

(b) 实验 2: 测试者认为给定音频与干净音频相同的比例

(c) 实验 3: 测试者认为给定音频与干净音频的说话人相同的比例

图 3 人类听觉测试

© 中国科学院软件研究所 http://www.jos.org.cn

实验 2: 判断给定的两个音频是否相同. 参与者同时听取一个未知音频 (与实验 1 来源相同) 及其对应的干净音频, 并判断两者是否为同一段音频. 结果如图 3(b) 所示, 实验结果展示了被测试者认为未知音频与干净音频相同的比例. 作为参考基准, 当两个完全相同的干净音频进行比较时, 有 84% 的参与者认为它们相同. 在其他方法生成的对抗音频中, Inaudible 表现最佳, 有 50% 的测试者认为其与干净音频相同; 而 SMACK 方法仅有 30% 的测试者判断为相同, 这主要源于其对音频韵律的修改导致语速和时长产生可感知变化, 从而容易被识别出差异. QFA2SR 的表现相对较弱, 仅有 26% 的参与者认为其与原始音频相同. 相比之下, SIAttack-c 和 SIAttack-s 分别获得了 60% 和 52% 的测试中判断为相同, 显著优于其他对比方法. 这一结果表明, 本文方法生成的对抗音频在听觉层面与原始音频保持了较高相似性, 具备更优的隐蔽能力.

实验 3: 判断给定的两个音频的说话人是否相同. 参与者同时听取两个音频样本, 并需判断它们是否由同一说话人录制. 本实验旨在验证在说话人信息层面添加扰动是否会影响人类对说话人身份的感知. 实验结果如图 3(c) 所示. 作为参考基准, 当参与者聆听两个未经扰动的干净音频时, 100% 的测试者正确判断其为同一说话人. 对于大多数攻击方法, 尽管在音频中引入了不同程度的扰动, 仍有约 90% 的参与者认为说话人相同, 表明人类听觉对身份特征的感知具有一定鲁棒性. 约 10% 的判断差异可能源于噪声干扰对听感造成的细微影响. Kenansville 方法的表现相对较低, 这很可能与其较差的音频质量有关, 失真程度较高导致说话人辨识困难. 相比之下, SIAttack-c 和 SIAttack-s 分别获得了 98% 和 96% 的说话人相同判断率, 显著优于其他对比方法. 这一结果说明, 无论是对内容信息还是说话人信息施加扰动, SIAttack 均能有效维持原始说话人的听觉身份特征, 未对人类的说话人辨识产生实质性干扰, 体现出其在保持音频身份特征方面的优越鲁棒性.

综合上述 3 项人类听觉实验结果可以看出, 与 QFA2SR 方法相比, SIAttack 所生成的对抗音频在保持高迁移性的同时, 具备更为优越的听觉质量. 此外, 相较于 Inaudible、SMACK 等专注于提升隐蔽性的方法, SIAttack 在维持与原始音频的全局相似度方面表现更加出色, 实现了攻击性能与隐蔽性之间的平衡.

3.6 防御性分析

本节将对各种攻击方法生成的音频在存在各种防御措施时的攻击性能进行评估. 在本节实验中, 考虑了 10 种较为常见的防御措施, 具体如下: 4 种时域防御方法, 分别为量化 (quantization, QT)、音频扰动 (audio turbulence, AT)、平均平滑 (average smoothing, AS) 以及中值平滑 (median smoothing, MS); 2 种频域防御方法, 即降采样 (down sampling, DS) 和低通滤波 (low pass filter, LPF); 1 种音频压缩方法: MP3-C; 3 种特征域防御方法, 分别是特征压缩 (feature compression, FeCo-d)[32]、变分自编码器防御 (variational autoencoder defense, VAE)[49]以及并行波形生成 (parallel waveGAN, PWG)[49]. 在本次实验中, 测试音频选取自表 6 设置下的对抗音频. 表 9 和表 10 则分别呈现了在部署各类防御措施时, 不同方法所生成的对抗音频在 OSI-u 和 OSI-t 两个任务上的攻击成功率. 其中, 表中的“None”表示未采取任何防御措施.

表 9 在部署防御措施时, 各种方法在 OSI-u 任务上的攻击成功率 (%)

Attack	None	QT	AT	AS	MS	DS	LPF	MP3-C	FeCo-d	VAE	PWG
Clean	0.0	3.8	3.8	2.3	5.2	1.7	2.4	3.8	6.2	4.2	5.8
FGSM	55.7	15.4	14.8	42.7	36.3	44.6	44.6	45.3	18.1	13.2	13.2
PGD	85.7	35.4	28.3	44.3	66.2	83.6	85.7	85.7	85.7	35.4	35.4
CW2	87.7	25.2	24.6	5.6	34.4	36.9	26.9	3.2	4.3	25.2	25.2
CWinf	85.6	35.4	28.3	44.3	66.2	83.6	85.3	85.3	3.9	35.4	35.4
Inaudible	82.5	25.2	24.6	5.6	34.4	36.9	26.9	3.2	4.4	25.2	25.2
Kenansville	76.3	37.3	34.6	57.7	42.1	66.7	66.7	66.3	64.3	53.2	53.2
QFA2SR	100.0	85.7	89.5	88.5	89.3	92.4	93.5	89.3	80.7	65.3	63.4
SirenAttack	80.4	11.3	0.6	14.5	13.5	39.1	11.3	45.3	12.6	14.7	12.8
FakeBob	96.0	14.8	15.4	17.5	15.8	44.4	14.8	53.8	15.3	14.8	13.5
SMACK	100.0	67.8	66.5	88.3	82.2	84.3	89.5	85.4	83.2	67.8	67.8
SIAttack-c	100.0	88.9	88.0	87.1	85.8	85.5	86.5	85.4	83.2	78.9	71.9
SIAttack-s	100.0	**98.2**	**99.0**	**97.8**	**98.8**	**98.4**	**99.0**	**97.5**	**91.3**	**85.9**	**83.9**

注: 加粗数据表示最优结果

© 中国科学院软件研究所 http://www.jos.org.cn

表 10 在部署防御措施时, 各种方法在 OSI-t 任务上的攻击成功率 (%)

Attack	None	QT	AT	AS	MS	DS	LPF	MP3-C	FeCo-d	VAE	PWG
Clean	0.0	3.8	3.8	2.3	5.2	1.7	2.4	3.8	6.2	10.2	9.8
FGSM	12.0	3.3	3.2	9.3	7.9	9.7	9.7	9.9	3.9	2.9	2.9
PGD	40.7	16.8	13.3	20.8	31.2	39.3	40.3	40.3	40.3	16.7	16.7
CW2	43.7	12.6	12.2	2.8	17.0	18.2	13.3	1.6	2.1	12.5	12.5
CWinf	45.3	18.8	15.0	23.5	35.0	44.3	45.2	45.2	3.5	18.7	18.7
Inaudible	41.1	12.6	12.3	2.8	17.2	18.5	13.5	1.6	2.2	12.6	12.6
QFA2SR	90.7	**81.4**	80.5	83.2	79.4	71.8	89.8	80.8	61.9	42.7	45.6
SirenAttack	65.4	1.0	0.5	3.7	11.0	31.8	9.2	4.3	3.1	5.3	4.2
FakeBob	76.2	3.8	4.3	6.0	12.6	35.5	11.8	3.8	4.2	3.8	3.2
SMACK	82.5	57.5	66.2	74.6	69.5	71.3	75.7	72.2	70.3	47.3	47.3
SIAttack-c	86.0	47.8	73.3	78.1	75.5	60.7	79.5	85.4	83.2	52.9	51.4
SIAttack-s	**94.5**	54.1	**88.1**	**93.9**	**82.5**	**89.0**	**95.4**	**93.5**	**85.3**	**79.9**	**77.9**

注: 加粗数据表示最优结果

从防御机制的角度分析, QT、AT、AS、MS 和 DS 等方法主要通过平滑处理、下采样或噪声注入等方法, 旨在抑制或掩盖对抗音频中的高频扰动成分. 由于这类处理对低强度扰动较为敏感, 它们在应对 FGSM、CW2、Inaudible 和 FakeBob 等扰动幅度相对较小的攻击时表现出较好的防御效果. 然而, 对于 PGD、QFA2SR 等扰动强度更高的攻击方法, 此类防御的效果有限, 攻击成功率未出现显著下降. 相比之下, FeCo-d、VAE 与 PWG 等防御手段聚焦于特征域处理, 通过特征压缩或基于高维表示的重建操作以去除噪声干扰. 在此类更复杂的防御机制下, 大多数攻击方法的成功率均出现明显降低, 反映出特征层级防御对多种对抗扰动具有更广泛的抑制能力.

从攻击方法的设计原理来看, SMACK、Kenansville 和 SIAttack 等方法均未直接在原始音频波形上添加扰动, 而是分别通过调整音频韵律、修改频谱特征, 或针对内容信息与说话人信息等高维语义特征施加扰动来实现攻击. 由于这类扰动并不表现为传统的高频噪声形式, 使得它们能够有效规避多数以滤除高频对抗噪声为主要目标的防御机制. 因此, 在面临多种防御措施时, 这些方法仍能维持较高的攻击成功率, 表现出更强的鲁棒性.

与现有攻击方法相比, SIAttack 在 OSI-u 与 OSI-t 两种任务设定以及多种防御环境下均展现出更优的攻击性能及更低的效能损失. 然而需要指出的是, 在 OSI-t 任务中面对量化 (QT) 防御时, SIAttack 的鲁棒性出现明显下降. 这一现象主要源于 QT 方法通过量化操作改变了音频信号的原始特征分布, 而 OSI-t 任务对文本相关特征的稳定性要求较高, 量化所引入的分布偏移可能破坏对抗音频中关键语义特征的完整性, 从而削弱了攻击效果. 此外, 实验结果显示 SIAttack-s 在面对 VAE 与 PWG 等基于重建的防御方法时, 比 SIAttack-c 表现出更强的鲁棒性, 这可能是因为说话人信息的扰动在 VAE 和 PWG 防御过程中更难被捕捉和消除.

总体而言, 通过对多种防御环境下各攻击方法的成功率进行综合评估与比较, 可以得出结论: SIAttack-s 在面对各类防御措施时均展现出更高的鲁棒性. 这表明该方法对不同攻击场景与防御机制具备更强的适应性与抗干扰能力, 能够以更稳定和有效的方式实现其攻击目标.

4 讨 论

4.1 均匀分布参数 p 对迁移性的影响

在第 3 节的实验中, 均匀分布参数被设置为 $p = 0.05$. 为测试均匀分布参数 p 对 SIAttack 生成的对抗音频的迁移性的影响, 本节将 p 设置为 0、0.001、0.01、0.05、0.1 和 0.5 这 6 个数值, 其余设置不变. 在不同 p 设置下, 首先使用 XV_PLDA、ECAPA 和 ResNet 这 3 个替代模型生成对抗音频, 然后将这些对抗音频迁移至 IV_PLDA 模型并测试其攻击成功率.

表 11 展示了不同 p 设置下迁移性的变化情况. 可知, p 的设置对 SIAttack-c 和 SIAttack-s 的迁移性均有影响. 当 $p = 0$ 时, 对抗音频的攻击成功率分别为 80.3% 和 89.3%. 在合理范围内增大 p 值有助于提升迁移成功率, 这是

© 中国科学院软件研究所 http://www.jos.org.cn

因为适度引入随机性有助于避免对抗音频在优化过程中陷入局部最优, 从而增强其泛化能力. 而在 $p > 0.05$ 时, 继续增加 p 值反而导致迁移性能下降, 说明过强的扰动干扰会增加对抗音频的生成难度, 削弱其有效性.

表 11 不同 p 设置下的迁移性 (%)

Attack	p=0	p=0.001	p=0.01	p=0.05	p=0.1	p=0.5
SIAttack-c	80.3	81.2	83.6	86.0	84.25	80.24
SIAttack-s	89.3	90.2	92.2	94.5	91.3	88.7

4.2 扰动阈值 ε 对音频隐蔽性的影响

在第 3 节的实验中, 阈值 ε 的大小根据第 2.5 节提出的阈值评估算法确定. 本节将评估阈值 ε 设置的有效性: 1) 分别在设置阈值 ε (由阈值评估算法确定) 和不设置阈值 (即阈值无穷大) 两种情况下, 生成相对应的对抗音频, 并从中挑选 6 名说话人的音频 (说话人编号分别为 P226、P227、P228、P229、P230 和 P232); 2) 向 50 名测试者提供一个未知音频 (来自以上的 6 名说话人的对抗音频或干净音频), 及 8 个参考音频 (来自以上 6 名说话人以及 2 名干扰说话人的干净音频), 并让测试者从参考音频中找出与未知音频声音最相似的音频.

表 12 展示了测试者对给定的未知音频判断正确的比例, Clean 表示当提供的未知音频为干净音频时, 测试者判断正确的比例. 例如, 当提供的未知音频为说话人编号为 P229 的干净音频时, 78% 的测试者认为提供的音频的声音与参考音频中 P229 的声音相同, 而剩下的 22% 的测试者则做出其他判断. SIAttack-s (无阈值) 和 SIAttack-s (有阈值) 分别展示了在没有设置阈值和使用扰动阈值算法设置阈值时, 生成的音频被测试者判断正确的比例. 由表 12 可见, SIAttack-s (有阈值) 生成的音频的正确识别率与 Clean 音频的差别在 4% 之内, 说明这些音频与 Clean 音频在听觉上相差不大. 相比之下, SIAttack-s (无阈值) 生成的音频的正确识别率较低, 说明此类音频与 Clean 音频在听觉上存在较大差异. 以上实验结果证明了扰动阈值评估算法的有效性.

表 12 测试者对给定的音频的说话人判断正确的比例 (%)

未知音频	P226	P227	P228	P229	P230	P232
Clean	68	68	76	78	80	50
SIAttack-s (无阈值) 生成的音频	56	58	66	68	68	48
SIAttack-s (有阈值) 生成的音频	66	66	76	76	78	50

注: P226、P228、P232为男性说话者; P228、P229、P230为女性说话者

4.3 不同置信度设置下的攻击效率

在第 3 节的实验中, 置信度设置为 $\kappa = 3/0.1$ (即在对 IV_PLDA 和 XV_PLDA 模型的攻击时设置为 3, 对其余模型攻击时设置为 0.1). 本节为测试不同置信度对攻击效果的影响, 分别将 κ 设置为 0/0.0、3/0.1 和 5/0.3. 随后, 在黑盒设置下, 针对不同迭代步数的情况, 对 SIAttack-s 在 4 个目标模型的 OSI-t 任务进行攻击效率测试. 图 4 展示了不同置信度下的攻击效率.

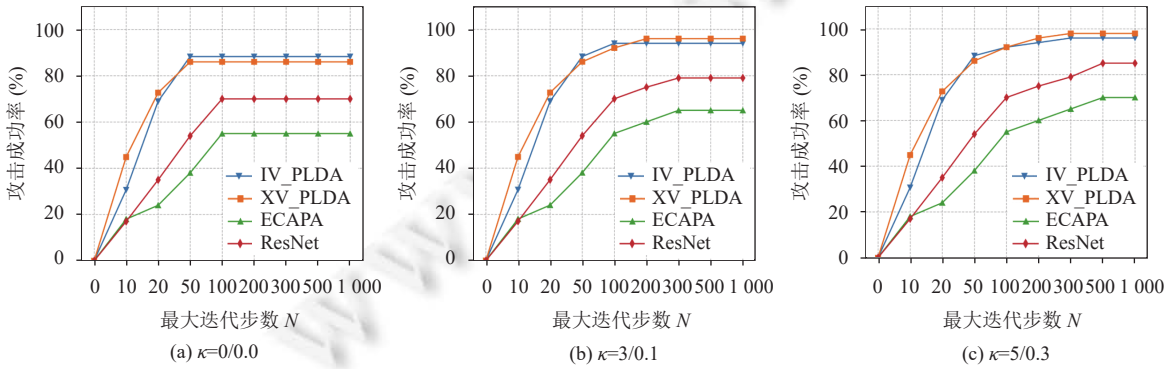


图 4 不同置信度下的攻击效率

由图 4 可知, 不同模型所需的迭代步数存在差异. 具体而言, 针对 IV_PLDA 和 XV_PLDA 模型发起攻击的难度相对较低, 其所需的迭代步数较少; 而针对 ResNet 和 ECAPA 模型的攻击难度则相对较高, 需要更多的迭代次数才能达到预期效果. 此外, 不同置信度水平下的攻击效率呈现出明显的差别. 置信度越高, 需要进行的迭代次数就越多, 同时生成的对抗音频的攻击成功率也相应越高. 在 $\kappa = 3/0.1$ 的设置情形下, 平均需要 300 次迭代才能够完成攻击. 虽然提高置信度在一定程度上能够提升攻击效果, 但这也会导致所需的迭代次数增多, 进而使得资源消耗大幅增加.

5 总 结

本文提出了一种基于说话人信息的音频对抗攻击方法——SIAttack. 该方法首先对原始音频进行信息解耦, 随后在解耦得到的说话人信息上通过迭代优化施加合理范围内的对抗扰动, 并重构生成对抗音频, 以有效攻击说话人识别系统. 大量实验结果表明, SIAttack 生成的对抗音频能够成功误导当前主流说话人识别模型. 与现有对抗攻击方法相比, SIAttack 所生成的样本在迁移性、隐蔽性与鲁棒性这 3 个方面均表现出更优的性能. 本研究揭示了在音频深层语义信息层面实施对抗攻击的可行性, 为音频对抗攻击领域的研究提供了新的方向.

References

[1] Kinnunen T, Li HZ. An overview of text-independent speaker recognition: From features to supervectors. Speech Communication, 2010, 52(1): 12–40. [doi: 10.1016/j.specom.2009.08.009]

[2] Markowitz JA. Voice biometrics. Communications of the ACM, 2000, 43(9): 66–73. [doi: 10.1145/348941.348995]

[3] Singh S. Forensic and automatic speaker recognition system. Int'l Journal of Electrical and Computer Engineering (IJECE), 2018, 8(5): 2804–2811. [doi: 10.11591/ijece.v8i5.pp2804-2811]

[4] Gu ZQ, Hu WX, Zhang CJ, Lu H, Yin LH, Wang L. Gradient shielding: Towards understanding vulnerability of deep neural networks. IEEE Trans. on Network Science and Engineering, 2021, 8(2): 921–932. [doi: 10.1109/TNSE.2020.2996738]

[5] Huang WH, Dai YS, Fei JW, Huang FJ. New visible watermark protection mechanism based on information hiding. IEEE Trans. on Information Forensics and Security, 2025, 20: 7764–7776. [doi: 10.1109/TIFS.2025.3592572]

[6] Lu H, Jin CJ, Helu XH, Zhu CS, Guizani N, Tian ZH. AutoD: Intelligent blockchain application unpacking based on JNI layer deception call. IEEE Network, 2021, 35(2): 215–221. [doi: 10.1109/MNET.011.2000467]

[7] Kreuk F, Adi Y, Cisse M, Keshet J. Fooling end-to-end speaker verification with adversarial examples. In: Proc. of the 2018 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018. 1962–1966. [doi: 10.1109/ICASSP.2018.8462693]

[8] Li ZH, Shi C, Xie Y, Liu J, Yuan B, Chen YY. Practical adversarial attacks against speaker recognition systems. In: Proc. of the 21st Int'l Workshop on Mobile Computing Systems and Applications. Austin: ACM, 2020. 9–14. [doi: 10.1145/3376897.3377856]

[9] Xie Y, Shi C, Li ZH, Liu J, Chen YY, Yuan B. Real-time, universal, and robust adversarial attacks against speaker recognition systems. In: Proc. of the 2020 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020. 1738–1742. [doi: 10.1109/ICASSP40776.2020.9053747]

[10] Chen GK, Chenb S, Fan LL, Du XN, Zhao Z, Song F, Liu Y. Who is real Bob? Adversarial attacks on speaker recognition systems. In: Proc. of the 2021 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2021. 694–711. [doi: 10.1109/SP40001.2021.00004]

[11] Qin Y, Carlini N, Cottrell G, Goodfellow I, Raffel C. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 5231–5240.

[12] Wang Q, Guo PC, Xie L. Inaudible adversarial perturbations for targeted attack in speaker recognition. In: Proc. of the 21st Annual Conf. of the Int'l Speech Communication Association. Shanghai: ISCA, 2020. 4228–4232.

[13] Zhang GM, Yan C, Ji XY, Zhang TC, Zhang TM, Xu WY. DolphinAttack: Inaudible voice commands. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. Dallas: ACM, 2017. 103–117. [doi: 10.1145/3133956.3134052]

[14] Chen T, Shangguan LF, Li ZJ, Jamieson K. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. In: Proc. of the 27th Annual Network and Distributed System Security Symp. San Diego: The Internet Society, 2020.

[15] Yuan XJ, Chen YX, Zhao Y, Long YH, Liu XK, Chen K, Zhang SZ, Huang HQ, Wang XF, Gunter CA. CommanderSong: A systematic approach for practical adversarial voice recognition. In: Proc. of the 27th USENIX Security Symp. Baltimore: USENIX Association, 2018. 49–64.

[16] Li ZH, Wu Y, Liu J, Chen YY, Yuan B. AdvPulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond

© 中国科学院软件研究所 http://www.jos.org.cn

perturbations. In: Proc. of the 2020 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2020. 1121–1134. [doi: 10.1145/3372297.3423348]

[17] Lin YQ, Abdulla WH. Principles of psychoacoustics. In: Lin YQ, Abdulla WH, eds. Audio Watermark: A Comprehensive Foundation Using Matlab. Cham: Springer, 2015. 15–49. [doi: 10.1007/978-3-319-07974-5_2]

[18] Xie CH, Zhang ZS, Zhou YY, Bai S, Wang JY, Ren Z, Yuille AL. Improving transferability of adversarial examples with input diversity. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2725–2734. [doi: 10.1109/CVPR.2019.00284]

[19] Chen GK, Zhang YD, Zhao Z, Song F. QFA2SR: Query-free adversarial transfer attacks to speaker recognition systems. In: Proc. of the 32nd USENIX Security Symp. Anaheim: USENIX Association, 2023. 2437–2454.

[20] Zhang Y, Li HW, Xu GW, Luo XZ, Dong GS. Generating audio adversarial examples with ensemble substituted models. In: Proc. of the 2021 IEEE Int'l Conf. on Communications. Montreal: IEEE, 2021. 1–6. [doi: 10.1109/ICC42927.2021.9500431]

[21] Xue M, Peng K, Gong XL, Zhang Q, Chen YJ, Li RT. Echo: Reverberation-based fast black-box adversarial attacks on intelligent audio systems. Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2023, 7(3): 137. [doi: 10.1145/3610874]

[22] Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 2000, 10(1–3): 19–41. [doi: 10.1006/dspr.1999.0361]

[23] Dehak N, Dehak R, Kenny P, Brümmer N, Ouellet P, Dumouchel P. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Proc. of the 10th Annual Conf. of the Int'l Speech Communication Association. Brighton: ISCA, 2009. 1559–1562.

[24] Variani E, Lei X, McDermott E, Moreno IL, Gonzalez-Dominguez J. Deep neural networks for small footprint text-dependent speaker verification. In: Proc. of the 2014 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Florence: IEEE, 2014. 4052–4056. [doi: 10.1109/ICASSP.2014.6854363]

[25] Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S. X-vectors: Robust DNN embeddings for speaker recognition. In: Proc. of the 2018 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018. 5329–5333. [doi: 10.1109/ICASSP.2018.8461375]

[26] Chen GK, Zhao Z, Song F, Chen S, Fan LL, Liu Y. AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems. IEEE Trans. on Dependable and Secure Computing, 2022: 1–17. [doi: 10.1109/TDSC.2022.3189397]

[27] Li JY, Tu WP, Xiao L. FreeVC: Towards high-quality text-free one-shot voice conversion. In: Proc. of the 2023 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023. 1–5. [doi: 10.1109/ICASSP49357.2023.10095191]

[28] Liu SX, Cao YW, Wang DS, Wu XX, Liu XY, Meng HL. Any-to-many voice conversion with location-relative sequence-to-sequence modeling. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2021, 29: 1717–1728. [doi: 10.1109/TASLP.2021.3076867]

[29] Kim J, Kong J, Son J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 5530–5540.

[30] Panayotov V, Chen GG, Povey D, Khudanpur S. LibriSpeech: An ASR corpus based on public domain audio books. In: Proc. of the 2015 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. South Brisbane: IEEE, 2015. 5206–5210. [doi: 10.1109/ICASSP.2015.7178964]

[31] Wu ZZ, Khodabakhsh A, Demiroglu C, Yamagishi J, Saito D, Toda T, King S. SAS: A speaker verification spoofing database containing diverse attacks. In: Proc. of the 2015 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. South Brisbane: IEEE, 2015. 4440–4444. [doi: 10.1109/ICASSP.2015.7178810]

[32] Chen GK, Zhao Z, Song F, Chen S, Fan LL, Wang F, Wang J. Towards understanding and mitigating audio adversarial examples for speaker recognition. IEEE Trans. on Dependable and Secure Computing, 2023, 20(5): 3970–3987. [doi: 10.1109/TDSC.2022.3220673]

[33] KALDI. VoxCeleb Models. 2022. https://kaldi-asr.org/models/m7

[34] KALDI. SITW Models. 2022. https://kaldi-asr.org/models/m8

[35] Desplanques B, Thienpondt J, Demuynck K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: Proc. of the 21st Annual Conf. of the Int'l Speech Communication Association. Shanghai: ISCA, 2020. 3830–3834.

[36] Faundez-Zanuy M, Monte-Moreno E. State-of-the-art in speaker recognition. IEEE Aerospace and Electronic Systems Magazine, 2005, 20(5): 7–12. [doi: 10.1109/MAES.2005.1432568]

[37] Tencent Cloud. Tencent automatic speech recognition. 2025 (in Chinese). https://cloud.tencent.com/product/asr

[38] IFlytek. IFlytek voiceprint recognition. 2025 (in Chinese). https://www.xfyun.cn/services/voiceprint-recognition

[39] Unisound. Unisound voiceprint recognition. 2025 (in Chinese). https://ai-poc.hivoice.cn/voiceprint-recognition

© 中国科学院软件研究所 http://www.jos.org.cn

[40] Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. In: Proc. of the 2001 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing. Salt Lake City: IEEE, 2001. 749–752. [doi: 10.1109/ICASSP.2001.941023]

[41] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: OpenReview.net, 2015.

[42] Mądry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018. 1050.

[43] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text. In: Proc. of the 2018 IEEE Security and Privacy Workshops. San Francisco: IEEE, 2018. 1–7. [doi: 10.1109/SPW.2018.00009]

[44] Du TY, Ji SL, Li JF, Gu QC, Wang T, Beyah R. SirenAttack: Generating adversarial audio for end-to-end acoustic systems. In: Proc. of the 15th ACM Asia Conf. on Computer and Communications Security. Taipei: ACM, 2020. 357–369. [doi: 10.1145/3320269.3384733]

[45] Abdullah H, Rahman MS, Garcia W, Warren K, Yadav AS, Shrimpton T, Traynor P. Hear "no evil", see "kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems. In: Proc. of the 2021 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2021. 712–729. [doi: 10.1109/SP40001.2021.00009]

[46] Yu ZY, Chang Y, Zhang N, Xiao CW. SMACK: Semantically meaningful adversarial audio attack. In: Proc. of the 32nd USENIX Security Symp. Anaheim: USENIX Association, 2023. 3799–3816.

[47] Wang XB, Hou R, Zhao BY, Yuan FK, Zhang J, Meng D, Qian XH. DNNGuard: An elastic heterogeneous DNN accelerator architecture against adversarial attacks. In: Proc. of the 25th Int'l Conf. on Architectural Support for Programming Languages and Operating Systems. Lausanne: ACM, 2020. 19–34. [doi: 10.1145/3373376.3378532]

[48] Credamo. 2025 (in Chinese). https://www.credamo.com/home.html#/

[49] Joshi S, Villalba J, Żelasko P, Moro-Velázquez L, Dehak N. Study of pre-processing defenses against adversarial attacks on state-of-the-art speaker recognition systems. IEEE Trans. on Information Forensics and Security, 2021, 16: 4811–4826. [doi: 10.1109/TIFS.2021.3116438]

附中文参考文献

[37] 腾讯. 腾讯云语音识别. 2025. https://cloud.tencent.com/product/asr

[38] 讯飞. 声纹识别. 2025. https://www.xfyun.cn/services/voiceprint-recognition

[39] 云知声. 声纹识别. 2025. https://ai-poc.hivoice.cn/voiceprint-recognition

[48] 见数. 2025. https://www.credamo.com/home.html#/

作者简介

陈家源, 硕士生, 主要研究领域为语音对抗攻击和防御.

黄文弘, 博士生, 主要研究领域为人工智能安全.

黄方军, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为人工智能安全, 多媒体内容安全.

© 中国科学院软件研究所 http://www.jos.org.cn