自动说话人识别中的模型反演攻击简介

Karla Pizzi[*, 1, 2], *Franziska Boenisch*[*, 1], *Ugur Sahin*[*, 1, 2], *Konstantin B¨ottinger*[1]

[1]Fraunhofer AISEC，德国； [2]慕尼黑工业大学，德国

[firstname.lastname]@aisec.fraunhofer.de

arXiv:2301.03206v1 [cs.SD] 9 Jan 2023

摘要

模型反演（MI）攻击可以重构机器学习（ML）模型训练数据中各类别的平均表示。研究表明，在每个类别对应不同个体的场景中，例如人脸分类器，这会构成严重的隐私风险。在本研究中，我们探索了模型反演的一种新应用：从说话人识别系统中提取说话人的语音。我们提出了一种方法，可以（1）从训练好的机器学习模型中重构音频样本；（2）提取中间语音特征表示，从而深入了解说话人的生物特征。

因此，我们提出了对模型反演攻击的一种扩展，并将其称为滑动模型反演。我们的滑动模型反演通过迭代反演音频样本的重叠分块来扩展标准模型反演，从而利用音频数据的序列特性提升反演效果。我们表明，可以利用反演音频数据生成伪造音频样本来冒充说话人，并代表其执行高度安全系统中的语音保护命令。据我们所知，我们的工作首次将模型反演攻击扩展到音频数据；研究结果揭示了在此场景下提取生物特征数据所带来的安全风险。**索引词**：说话人识别、模型反演、隐私

1. 引言

对音频数据的隐私分析表明，口音、节奏或语音的声学属性等语音参数，天然携带有关说话人的生物特征信息，例如年龄、性别、身体健康状况和地理来源 [1]。因此，说话人识别中的机器学习（ML）模型不泄露其训练数据的信息至关重要。然而，近期研究 [2, 3, 4]表明，机器学习模型普遍容易受到隐私攻击。其中一种攻击是模型反演（MI） [2]，攻击者可借此获取目标模型训练数据中各个类别的抽象表示。由于说话人识别系统将每个人都视为一个类别，模型反演攻击可能导致严重的隐私泄露 [2]。迄今为止，模型反演攻击对说话人识别系统和音频数据是否可行尚未经过验证，因此，能否恶意获取说话人信息这一问题仍悬而未决。

我们首次展示了如何针对音频数据调整并应用模型反演攻击。我们以 SincNet 为攻击目标；它是一种基于神经网络（NN）的先进说话人识别模型 [5, 6]。我们展示了，模型反演攻击能够从训练后的目标模型中推断出完整的音频样本，以及作为说话人语音特征中间表征的 d 向量。此外，我们提出了滑动模型反演，这是一种新颖的标准模型反演攻击形式，它利用音频数据的序列处理特性来提高反演成功率。采用标准模型反演时，目标模型能够成功识别出多达 54% 的反演音频样本，并将其归入正确的说话人类别；而采用我们新颖的攻击，准确率最高可达 90%。此外，我们的滑动模型反演能够缩小原始样本与反演样本在 d 向量表征中的距离，从而生成保真度更高的反演结果。对于直接反演 d 向量，我们的实验表明，即使采用标准模型反演，识别成功率也能达到 100%。这些结果凸显了说话人识别模型容易受到隐私攻击。作为概念验证，为展示我们的模型反演可作为进一步攻击说话人识别系统的切入点，我们还探索了将反演音频样本用作深度伪造音频生成的输入。这类深度伪造音频可用于通过任意语音样本欺骗语音识别系统，或冒充遭受攻击的说话人执行任何语音命令。尽管作者进行的非正式评估表明，我们生成的深度伪造音频无法完美欺骗人类听者，但它们说明，隐私攻击不仅能够泄露模型训练所用个体的敏感信息，还可能严重威胁依赖语音生物特征的系统安全。我们的贡献总结如下：

- 我们成功地对说话人识别模型实施成员推断攻击，以反演完整的音频样本和 d 向量，并通过实验评估哪种随机初始化最适合作为针对音频数据的成员推断攻击输入。
- 我们提出了一种新颖的 滑动互信息，它利用了序列化和分块音频处理的特性。
- 我们展示了基于推断出的音频样本生成深度伪造音频的可行性。

2. 背景与相关工作

以下章节将介绍说话人识别系统以及针对其隐私的攻击的背景信息。

说话人识别。 本文采用一种基于 SincNet [5] 的文本无关说话人识别系统。该系统利用神经网络提取语音特征，并将其转换为所谓的 d 向量，再在其上添加一个分类层。系统的输入为原始音频波形，输出则是所有可能类别（即说话人）的逐类概率分数。总体而言，该系统由三个子模型组成（见图1）：1）SincNet，一种类似带通滤波器的卷积神经网络；2）用于计算 d 向量的多层感知机（MLP） [7]；以及 3）用于计算每位说话人对应概率的全连接层。据报告，该系统在 TIMIT [8] 测试数据集上取得了 $5.772 \cdot 10^{-3}$ 的分类错误率[1]（以句子为单位测量）。在当前版本中，SincNet 未提供任何专门的隐私保护机制。

*作者贡献相同。

[1]参见 https://pythonlang.dev/repo/ mravanelli-sincnet/。

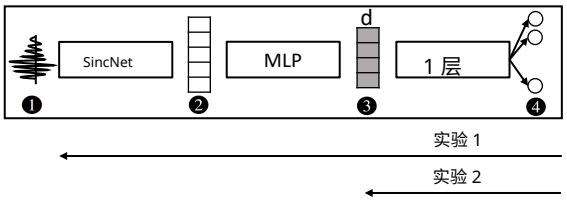

图1：说话人识别模型。说话人识别模型及其三个子模型：*SincNet* 获取 原始音频输入并生成 特征。这些特征输入多层感知机后，会生成 *d* 向量。单层网络对其进行 分类。实验 *1* 中，我们反演完整音频样本；实验 *2* 中，我们反演 *d* 向量。

算法 1标准模型反演攻击 [2]

函数 模型反演（输入向量 x_0，目标类别 t，迭代次数 α，耐心值 β，最低成本阈值γ，梯度下降的学习率 λ）

对于 $i \leftarrow 1$，...， α **执行** $x_i \leftarrow x_{i-1} - \lambda \cdot \nabla c$ (x_{i-1}) **如果** $c(x_i) \geq \max(c(x_{i-1})$ ...， $c(x_{i-\beta}))$ **则跳出 如果** $c(x_i) \leq \gamma$ **则跳出 返回** [$\text{argmin}_{x_i} c(x_i)$， $\min_{x_i} c(x_i)$]

说话人识别中的隐私。 ISO/IEC 标准 24745:2011 提出了确保个人隐私的三项一般要求：受保护数据的不可逆性、可更新性和不可关联性 [9]。为实现这一目标，针对说话人识别系统，人们已提出并讨论了多种解决方案 [9]。然而，在说话人识别模型中，任何旨在保护说话人隐私的匿名化 措施，都与系统根据说话人的个体特征识别说话人的目的直接冲突。

成员推断攻击。在 MI 攻击中 [2], ，攻击者利用机器学习模型的预测置信度，反演各个训练类别（见算法 1）。更正式地说，MI 攻击可以表述如下：设 f为受到攻击的目标模型。该模型经过训练，可将一个 n维输入数据点x 映射为一个 m维向量 p，用于表示各类别的概率，使得 f： $x \mapsto p$，其中 $\mathbb{R}^n \to [0,1]^m$。为了反演模型，我们定义一个目标函数，以便使用梯度下降。该函数称为代价函数 c（x），它基本上用于衡量我们与希望重建的信息有多接近。我们设定 c（x） $= 1 - p_t$，其中 t 表示我们希望获取其相关信息的目标类别。从一个随机初始化的输入样本 x_0开始，我们计算其代价c（x_0）。有了这一结果，我们便以学习率 λ 运行梯度下降算法 α 次迭代，从而修改原始输入。目标是最小化特定类别 [2],的代价，使得到的数据样本能够代表该类别。在说话人识别中，每位说话人都代表一个独立类别。因此，模型反演可以揭示训练数据集中每个个体的数据表示。具体而言，这些数据可能包含生物特征以及副语言特征。

算法 2 滑动模型反演攻击

函数 SMI(目标类别 t，长度 l，步长 s，窗口大小 w，α，β，γ，λ如算法 1 所示)
反演[0,。 。 。 ，l] $\leftarrow [\mathcal{N}(\mu, \sigma^2)]^l$
对于 $k \leftarrow 0$，.. .，（$l - s$）**按照步长** s **执行**
$x \leftarrow$ 反演[k： （$k + w$）]
对于 $i \leftarrow 1$, ˚ ˚ ˚, α **do**
$x_i \leftarrow x_{i-1} - \lambda \cdot \nabla c(x_{i-1})$
如果 $c(x_i) \geq$最大值($c(x_{i-1})$ .. ., $c(x_{i-\beta})$) **则**
中断
如果 $c(x_i) \leq \gamma$ **则**
中断
反演[k : ($k + w$)] $\leftarrow \text{argmin}_x\, c(x)$
返回 反演[$\frac{w}{2}$： ($t - \frac{w}{2}$)] −

3. 滑动式 MI 攻击

本节介绍我们提出的新型滑动式 MI 攻击。它将标准MI（见算法 1）扩展到按序列和分块处理的数据，例如音频数据。我们的滑动式 ML 并非使用 MI 分别反演每个数据块，而是迭代地反演重叠分块。这样一来，MI 的部分输入已经过反演，因此其波形与真实语音数据相似。由此，MI 能够更成功地将其反演为原始语音数据的表征。

我们的滑动互信息包括以下步骤：(1) 我们对随机初始化的输入向量的第一个窗口进行反演。请注意，不同的随机初始化会产生质量不同的反演结果。作为我们的第一个主要实验，我们测试了几种不同类型的初始化设置，详见第 4.1 节。(2) 然后，我们使用所得的反演数据替换输入向量的第一个窗口。(3) 接下来，我们迭代计算后续输入窗口的反演数据。该输入的第一部分由先前的反演数据组成，第二部分则来自随机输入向量。请注意，滑动窗口的重叠量决定了在 MI攻击期间用于计算反演的先前反演数据与随机初始化的输入向量所占的比例。其值取决于我们的反演步长 。在我们的实验中，我们使用 500 个采样点的步长（约 30 毫秒）。由于在这种新方法中，更新依赖于先前反演的输出，因此我们无法使用并行化来加速。相反，步长决定了计算开销。增大步长值可以缩短计算时间。

有关我们新方法的可视化示意，请参见图 2。请注意，除标准MI的超参数外，我们还需要指定输入数据 l 的长度、步长 s 以及窗口大小 w。 s 指定了连续反演之间的重叠部分，而 w 决定了被反演的数据块长度。对于每个数据块，反演均按照标准MI中的方式进行迭代。由于被反演向量的开头和结尾部分经过的迭代次数较少，我们会将返回的向量裁剪至窗口大小的一半。有关我们新型滑动式 ML 的形式化介绍，请参见算法 2。

4. 实验

我们进行了三项实验：第一项实验 与其他领域中的模型反演类似，即将随机输入向量反演回原始输入数据域。在第二项实验中，我们不反演整个神经网络，而只反演截至 d 向量的各层；这些 d 向量包含个体独特的语音特征——

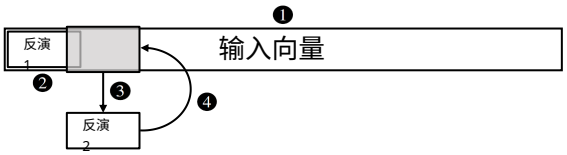

图 2：滑动互信息。 我们初始化一个随机输入向量。 从开头开始，我们基于该向量反演第一个窗口，并用反演得到的数据替换向量的第一部分。 对于之后的所有窗口，我们使用先前反演向量的部分内容，并用输入向量填充其余部分，以执行模型反演。 然后，我们迭代地用反演数据替换输入向量。

（见图1）。与反演完整音频样本相比，反演这些向量降低了攻击的计算成本。我们的第三项实验 表明，在说话人识别中，我们的模型反演攻击能够冒充个别说话人，并为他们合成语音样本。该欺骗攻击基于第一项实验中反演得到的音频样本。

我们使用在 TIMIT 数据集 [8], 上训练的 SincNet [5], 攻击基于神经网络的说话人识别系统，并使用 Ravanelli 等人 [5]提供的预训练模型。为了执行模型反演攻击，我们假设攻击者对白盒访问目标模型。这种情况例如可能出现在说话人识别模型部署在用户设备上时，例如用于生物特征识别。此外，攻击者需要知道其攻击目标个体的唯一标识符，以确定要反演训练数据中的哪个类别。例如，在 TIMIT 数据集的情况下，该标识符可以是姓名或某种假名化字符组合，例如“FGMB0”。

我们通过以下方式量化实验的成功程度：

- 正确分类的反演样本所占百分比。我们量化原始目标说话人模型在反演音频数据和反演 d 向量上的分类准确率。如果反演样本被分类为正确的原始说话人，则称其为“正确分类”。
- 原始 d 向量与重建的 d 向量之间的欧几里得距离。我们测量这两个 d 向量之间的欧几里得距离，以确定相应样本之间的相似度。

第一项指标使我们能够分析模型反演是否可视为在目标模型方面取得成功，也就是说，它回答了该模型能在多大程度上被欺骗。相比之下，第二项指标关注反演样本与原始样本的相似度。因此，它从人类听者的角度量化了相似度。

由于模型反演会生成训练类别的平均表示，我们将目标模型对按说话人平均的样本的分类准确率用作基线（训练数据和测试数据上的正确分类率分别为 97.84% 和 75.97%）。下面，我们先介绍整体实验设置，再详细说明各项实验。

4.1. 实验1：反演音频采样点

在第一个实验中，我们使用MI计算完整的反演音频样本，这些样本可用于在没有人类听众在场的情况下欺骗受到攻击的说话人识别模型。该实验旨在回答以下三个问题：
(1) 是否能够成功生成用于说话人识别的反演音频样本?
(2) 哪种随机初始化的
输入向量用于MI攻击时，能够产生最成功的反演音频样本（就被分类为原始说话人而言）？ (3) 与标准MI相比，我们的新型滑动互信息方法如何改进结果?

实验。 在所有实验中，音频数据块的长度均为 3200 个采样点（或 200 毫秒），滑动互信息的步长为 500 个采样点（约 30 毫秒）。我们评估了以下（随机）初始化：

- 纯输入：全零、全一和全负一；
- 噪声：白噪声、粉红噪声、布朗噪声、紫噪声和蓝噪声。我们使用 python-acoustics 库中预先实现的方法生成这些噪声，并应用 tanh 函数，以非线性方式将其转换到区间 $[-1, 1]$，且均值为零；
- 从分布中采样，例如均匀分布（范围为 $[0, 1]$ 或 $[-1, 1]$）、高斯分布（参数为 $\mu = 0$ 或 $\sigma = 0.2$）、拉普拉斯分布（参数为 $\mu = 0$ 或 $b = 0.07$）、贡贝尔分布（参数为 $\mu = 0$ 或 $\beta = 0.1$）和 von Mises 分布（参数为 $\mu = 0$ 或 $\kappa = 0.1$）；
- 来自另一数据集的采样点：Librispeech [10], 作为普通输入，或对 50、100、150 个输入向量取平均，或加入白噪声（$0.85 \cdot$ 输入向量 $+ 0.15 \cdot$ 噪声）。

对于 MI算法（见第 2 节）中的优化过程，我们优化两个参数，即最大迭代次数 α 和学习率 l。对于所有实验，设置 $\alpha = 1000$ 已足够。最优 l 取决于实验，并在结果中报告。在我们的评估中，对于每种输入及其设置和最优 l，我们报告以下指标：

- *MI*准确率：正确分类的反演采样点所占的百分比；
- # 正确分类的说话人数：正确分类的说话人数；
- 平均欧氏距离及标准差：在 d-vector空间中，反演样本与同一说话人的原始样本之间的平均欧氏距离，仅根据成功反演的音频样本计算。

对于 d-vector空间中的欧几里得距离，原始训练样本中说话人内距离的平均值可作为基线（$1.923 \cdot 10^{-1}$）。

结果。 *(1)* 是否有可能成功生成用于说话人识别的反演音频样本？ 通过观察反演样本的分布，我们发现它与原始数据的分布并不完全一致。由于反演样本听起来也与原始样本不同——它们不一定听起来像语音——因此无法欺骗人类听众。然而，结果表明，标准MI足以欺骗自动说话人识别系统的分类，准确率最高可达54.76%。采用我们提出的新型滑动互信息，分类准确率可大幅提升至90.48%。对于负责高度安全系统身份控制的说话人识别模型而言，这种基于反演数据得到的准确率远不可接受。与标准MI相比，我们提出的新型滑动互信息还降低了部分输入向量对应的反演样本与原始样本之间的欧几里得距离。然而，尽管距离有所缩短，重建语音样本听起来仍与原始说话人的声音相去甚远。

(2) 哪种随机初始化的输入向量用于模型反演攻击时，能生成最成功的反演音频样本（以分类为原始说话人的效果衡量）？ 我们还可以得出结论，并非所有输入向量都同样适合用于模型反演，以生成能够成功欺骗说话人模型的反演音频样本：普通输入向量在分类准确率方面的反演效果最差。我们推测，这是因为很难通过优化将恒定向量转换成类似语音的波形。对于音频数据而言，随机初始化的向量或本身已呈波形的数据似乎更适合作为模型反演的输入：使用白噪声和tanh 激活函数可获得最高的分类准确率。布朗噪声的表现最差，但其平均欧几里得距离相对较小。拉普拉斯分布的总体结果最佳。结果概览见表1。

（*3*）与标准*MI*相比，我们的新型滑动互信息方法如何提升结果？从表1的结果可以看出，滑动互信息的性能高于标准MI。使用标准MI时，目标模型在反演数据上的准确率为54%；而使用滑动互信息时，尽管结果取决于随机初始化，准确率仍可达到90%以上。

4.2. 实验2：对 d 向量进行部分MI反演

以往针对其他数据类型的应用中，只有将数据完全反演回原始输入域才有价值；但说话人识别的情况有所不同：d 向量是语音样本的特征表示，本身已包含重要的副语言信息，例如可用于生成伪造音频样本 [11]。反演简单的 d 向量而非完整音频样本，可以降低攻击的计算成本（因为无需按顺序逐一处理），并且可以通过标准MI攻击实现。因此，在第二个实验中，我们尝试对模型的中间层进行反演，旨在回答以下两个问题：（1）能否成功反演 d 向量？（2）哪个输入向量能生成最成功的反演 d 向量（以分类为原始说话人为准）？

实验。 我们通过移除网络中的 SincNet 和 MLP 部分，仅关注用于 d 向量反演的子模型3，来进行部分反演（见图1）。

结果。 （*1*）能否成功反演 *d* 向量？ 我们的研究结果表明，我们能够重建 d 向量，并使其被成功分类为原始说话人（目标模型在这些向量上的分类准确率达到 100%）。这一结果优于基线：在按说话人平均的采样点上测得的目标模型准确率为 97。84%。然而，即使对于表现最佳的输入（零），

$$0.84 \pm 0.028$$

的欧几里得距离也明显高于基线中的说话人内平均距离（$1.923 \cdot 10^{-1}$）。为评估反演后的 d 向量是否仍会泄露个体说话人的隐私，我们进行了主成分分析（PCA），并训练了一个二元分类器来预测个体的性别。我们在 TIMIT 测试数据集上拟合 PCA，并对反演后的 d 向量进行变换。结果如图3所示。结果表明，反演后的 d 向量会泄露性别隐私。

(2) 哪种输入向量能够产生最成功的反演 *d* 向量（就分类为原始说话人而言）？由于 d 向量和音频数据具有不同的属性，要成功进行反演，就需要使用不同的输入向量

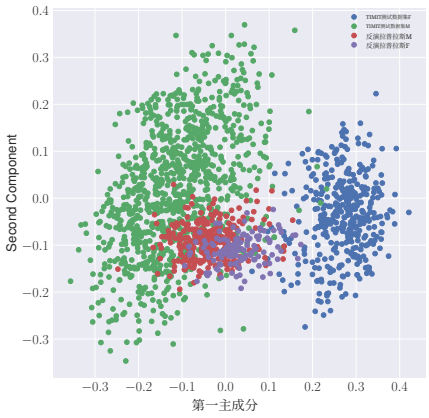

图3： ***d*** 向量的主成分分析。 在 *TIMIT* 测试数据集上拟合主成分分析（蓝色：女性；绿色：男性），并用于转换反演 *d* 向量（紫色：女性；红色：男性）。结果表明，反演 *d* 向量能够揭示个体的性别。

。声音噪声是音频样本的良好输入。然而，我们的观察表明，以声音噪声初始化输入向量并不能产生高质量的反演 d 向量表示。相反，我们发现，在反演 d 向量时，直接使用全零效果最好。

4.3. 实验3：创建深度伪造音频

尽管目标模型能够正确分类反演采样点，但这些采样点未必包含对人类听众有用的信息。在接下来的实验中，我们重点探讨以下问题：(1) 能否根据反演音频样本生成音频数据，使其在人类听者听来与原始说话人相似？这项实验可视为一项概念验证，用以展示我们的攻击可能给说话人识别系统带来的进一步安全风险。

实验。 为了生成深度伪造音频，我们采用了 [11] 的研究成果。其架构由三部分组成：（1）说话人编码器，（2）语音合成器，以及（3）声码器。说话人编码器用于从音频文件中提取 d 向量，以便在向量空间中表征音频样本。语音合成器利用这些 d 向量生成梅尔频谱图。最后，声码器获取梅尔频谱图，执行频域到时域转换。底层的语音合成器是 Tacotron 2 [12]，声码器则是 Wavenet [13]。我们将滑动互信息得到的反演音频样本和反演 d 向量作为该方法的输入。原则上，反演音频样本可以直接输入说话人编码器，用于生成深度伪造音频。然而，要使用我们反演得到的 d 向量（2048 维），必须将其转换为与深度伪造模型的说话人编码相匹配的形式，因为该模型要求 d 向量为 256 维。为此，我们训练一个多层感知机，将一个向量空间映射到另一个向量空间。该多层感知机有两个隐藏层，分别包含 1024 和 512 个神经元，输出层包含 256 个神经元。前两层使用 tanh 激活函数。我们将音频文件分别输入我们自己的说话人编码器和深度伪造模型的说话人编码器，以此构建用于该转换的训练集。我们得到的编码作为输入，而对方得到的编码则作为待学习的输出。

	样本类型	学习 R.	MI准确率	# 正确分类的说话人数	平均欧氏距离 ±标准差
plain inputs	ones	$1 \cdot 10^{-05}$	0.43%	2	0.800 ± 0.0416
		0.2	5.41%	25	0.696 ± 0.0663
	零	$1 \cdot 10^{-08}$	0.65%	3	0.792 ± 0.0377
		0.5	8.87%	41	0.720 ± 0.0682
dists	贡贝尔	0.01	54.76%	253	0.797 ± 0.0562
		0.01	89.83%	415	0.748 ± 0.0561
	拉普拉斯	0.005	54.33%	251	0.793 ± 0.0555
		0.005	90.48%	418	0.752 ± 0.0550
noise	白色双曲正切	0.2	54.11%	250	0.798 ± 0.0549
		0.2	88.31%	408	0.757 ± 0.0561
	布朗语料库	0.01	5.84%	27	0.676 ± 0.0938
		0.05	31.82%	147	0.681 ± 0.0679
Librispeech samples	LibriSpeech 样本	0.001	6.277%	29	0.751 ± 0.0768
		0.2	22.73%	105	0.696 ± 0.0558
	样本噪声 + LibriSpeech 样本	0.005	52.6%	243	0.779 ± 0.0673
		0.01	90.26%	417	0.757 ± 0.0595
	Librispeech 均值	0.001	33.55%	155	0.750 ± 0.0665
		0.01	56.06%	259	0.707 ± 0.0713
	样本噪声 + Libspeech 均值	0.005	45.89%	212	0.776 ± 0.0604
		0.01	80.30%	371	0.757 ± 0.0600

表1：结果 标准***MI***（灰色）和滑动互信息（黑色），根据***TIM***中***462***名说话人的反演音频样本计算得出 *IT* 数据集（*326*名男性，*1* *36*名女性）。滑动互信息优于标准*MI*（准确率更高，平均欧几里得距离更低。）

结果。（*1*）根据反演音频样本，能否生成让人类听者觉得像原始说话人的音频数据？如实验2所述，我们的 d 向量虽然被正确分类为原始说话人，但在欧几里得空间中与说话人的原始 d 向量相距甚远。生成样本听起来是否与原始样本相似，是一个语义问题，取决于上下文的敏感性。根据作者逐一检查的结果，基于 d 向量生成的深度伪造音频无法让人辨认出个别说话人的特征。不过，利用我们新提出的滑动互信息所得到的反演音频样本，我们成功生成了一些高质量的伪造音频样本，听起来与原始说话人相似。[2] 因此，攻击者只需利用从预训练神经网络反演得到的数据，就可能伪造某人的身份。我们预计，随着未来深度伪造生成系统日益复杂，这种情况会愈发普遍。

5. 对策与讨论

说话人识别系统在很大程度上依赖于学习每位说话人的个体特征，以完成其设计任务。因此，这些系统必然始终包含有关其训练所用说话人数据的信息。去除个体说话人特征会导致系统性能大幅下降。尤其是，假名化 [14] 和通用语音系统中采用的隐私保护方法（例如，[15, 16, 17, 18, 19, 20]）会使说话人识别无法实现其原本的用途。

防御MI攻击。作为替代方案，可以考虑旨在阻碍MI攻击的隐私保护方法。现有来自其他领域的大多数防御措施主要致力于抑制模型置信度分数或降低其效用。这可以通过向其中注入均匀噪声 [21], 、降低其精度 [2] 或降低其离散度 [22]来实现。后一种方法会降低输入数据与分数之间的相关性，从而使MI攻击更加不准确。出于类似目的，有研究指出，在训练损失函数中使用正则化也可以作为一种防御措施 [23]。此外，还可以采用面向硬件的解决方案，阻止攻击者访问模型参数，以降低MI攻击的成功率，或至少阻止其提取中间特征（参见我们的实验2） [24]。

差分隐私。早期研究通过实验证明，差分隐私（DP） [25] 可以降低 MI 攻击的成功率 [2]，但需要使用大量噪声，这反过来会严重降低模型的性能。后续研究表明，针对机器学习模型的差分隐私训练完全无法防止成员推断攻击 [26]，因为其目的是掩盖某个数据点是否存在于特定数据集中，而非保护数据类别的隐私。

局限性。迄今为止，MI攻击需要获取神经网络的置信度分数，且攻击的成功很大程度上取决于随机初始化。尤其是，许多说话人识别工具依赖余弦相似度 [27, 28]，因此需要更新算法。此外，伪造音频样本的质量受深度伪造生成方法的限制。因此，我们的攻击目前可能仍具有有限的实际影响。不过，随着新的、更强大的隐私攻击和深度伪造方法不断涌现，利用隐私攻击威胁说话人识别系统安全的风险将日益受到重视。因此，我们有必要提高相关意识，并将隐私与安全结合起来考虑和保护，而不是将二者割裂开来。

[2]原始音频数据、平均后和反演后的采样点，以及基于反演音频采样点生成的欺骗音频数据示例，可在 https://www.dropbox.com/sh/ge6xx90laqmru9b/AABUsS3p4EwaN0n7g4Rgq8rwa 获取。

6. 结论

在本研究中，我们首次成功地对音频数据实施了模型反演攻击。为此，我们提出了一种新颖的滑动式成员推断方法，利用音频数据的序列特性来改进反演。我们通过实验评估了该攻击对最先进的说话人识别系统的成功效果。结果表明，我们反演得到的音频样本可作为进一步攻击目标系统安全性的起点。由此，我们强调了在此类系统中实施充分隐私保护的重要性。

7. 致谢

本研究得到巴伐利亚州经济事务、区域发展和能源部的资助。作者感谢 J. Williams 对稿件提出的意见。

8. 参考文献

[1] J. L. Kr¨oger、O. H.-M. Lutz 和 P. Raschke，“语音与言语分析的隐私影响——通过推断进行信息披露”，载于 *IFIP*隐私与身份管理国际暑期学校。Springer，2019年，第242–258页。[2] M. Fredrikson、S. Jha 和 T. Ristenpart，“利用置信度信息的模型反演攻击及基本对策”，载于 第*22*届*ACM SIGSAC*计算机与通信安全会议论文集，2015年，第1322–1333页。[3] R. Shokri、M. Stronati、C. Song 和 V. Shmatikov，“针对机器学习模型的成员推断攻击”，载于 *2017*年*IEEE*安全与隐私研讨会（*SP*）。IEEE，2017年，第3–18页。[4] K. Ganju、Q. Wang、W. Yang、C. A. Gunter 和 N. Borisov，“使用置换不变表示对全连接神经网络实施属性推断攻击”，载于 *2018*年*ACM SIGSAC*计算机与通信安全会议论文集，2018年，第619–633页。[5] M. Ravanelli 和 Y. Bengio，“基于原始波形的说话人识别：SincNet”，载于 *2018*年*IEEE*口语技术研讨会（*SLT*）。IEEE，2018年，第1021–1028页。[6] M. Ravanelli、T. Parcollet、A. Rouhe、P. Plantinga、E. Rastorgueva、L. Lugosch、N. Dawalatabad、C. Ju-Chieh、A. Heba、F. Grondin、W. Aris、C.-F. Liao、S. Cornell、S.-L. Yeh、H. Na、Y. Gao、S.- W. Fu、C. Subakan、R. De Mori 和 Y. Bengio，“Speechbrain”，https://github.com/speechbrain/speechbrain，2021年。[7] E. Variani、X. Lei、E. McDermott、I. L. Moreno 和 J. Gonzalez-Dominguez，“用于小型化文本相关说话人验证的深度神经网络”，载于 *2014*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2014年，第4052–4056页。[8] J. S. Garofolo，“Timit声学语音连续语料库”，语言数据联盟，*1993*年，1993年。[9] A. Nautsch、A. Jim´enez、A. Treiber、J. Kolberg、C. Jassera nd、E. Kindt、H. Delgado、M. Todisco、M. A. Hmani、A. Mtibaa 等，“保护说话人与语音表征中的隐私”，*Computer Speech & Language*，第58卷，第441–480页，2019年。[10] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，“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* 声学、语音与信号处理国际会议（*ICASSP* ）。IEEE，2015年，第5206–5210页。[11] Y. Jia、Y. Zhang、R. J. Weiss、Q. Wang、J. Shen、F. Ren、Z. Chen、P. Nguyen、R. Pang、I. L. Moreno 等，“从说话人验证到多说话人文本转语音合成的迁移学习”，*arXiv*预印本 *arXiv:1806.04558*，2018年。[12] J. Shen、R. Pang、R. J. Weiss、M. Schuster、N. Jaitly、Z. Yang、Z. Chen、Y. Zhang、Y. Wang、R. Skerrv-Ryan 等，“通过以梅尔频谱图预测为条件的Wavenet实现自然文本转语音合成”，载于 *2018*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2018年，第4779–4783页。

[13] A. v. d. Oord、S. Dieleman、H. Zen、K. Simonyan、O. Vinyals、A. Graves、N. Kalchbrenner、A. Senior 和 K. Kavukcuoglu，“Wavenet：一种用于原始音频的生成模型”，arXiv预印本 arXiv:1609.03499，2016年。[14] P.-G. No´e、A. Nautsch、N. Evans、J. Patino、J.-F. Bonastre、N. Tomashenko 和 D. Matrouf，“迈向统一的语音假名化评估框架”，Computer Speech & Language，第72卷，第101299页，2022年。[15] R. Aloufi、H. Haddadi 和 D. Boyle，“平衡语音用户界面隐私与安全的串联框架”，arXiv预印本 arXiv:2107.10045，2021年。[16] Q. Jin、A. R. Toth、T. Schultz 和 A. W. Black，“Voice conver- gin：通过语音转换实现说话人去标识化”，载于2009年IEEE国际声学、语音与信号处理会议。IEEE，2009年，第3909–3912页。[17] B. M. L. Srivastava、N. Vauquier、M. Sahidullah、A. Bellet、M. Tommasi 和 E. Vincent，“针对知情攻击者评估基于语音转换的隐私保护”，载于 ICASSP 2020-2020年IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2020年，第2802–2806页。[18] B. M. L. Srivastava、N. Tomashenko、X. Wang、E. Vincent、J. Ya- magishi、M. Maouche、A. Bellet 和 M. Tommasi，“基于x-vector的说话人匿名化设计选择”，arXiv预印本 arXiv:2005.08601，2020年。[19] J. Qian、H. Du、J. Hou、L. Chen、T. Jung 和 X.-Y. Li，“Hide-behind：借助声纹不可克隆性与匿名性享受语音输入”，载于第16届ACM嵌入式网络传感器系统会议论文集，2018年，第82–94页。[20] B. M. L. Srivastava、A. Bellet、M. Tommasi 和 E. Vincent，“自动语音识别中的隐私保护对抗式表征学习：现实还是幻觉？”，arXiv预印本 arXiv:1911.04913，2019年。[21] A. Salem、A. Bhattacharya、M. Backes、M. Fritz 和 Y. Zhang，“Updates-leak：在线学习中的数据集推断与重构攻击”，载于第29届 {USENIX} 安全研讨会（{USENIX}Security 20），2020年，第1291–1308页。[22] Z. Yang、B. Shao、B. Xuan、E.-C. Chang 和 F. Zhang，“通过预测净化防御模型反演攻击和成员推断攻击”，arXiv预印本 arXiv:2005.03915，2020年。[23] T. Wang、Y. Zhang 和 R. Jia，“通过互信息正则化提升对模型反演攻击的鲁棒性”，arXiv预印本 arXiv:2009.05241，2020年。[24] Q. Xu、M. T. Arafin 和 G. Qu，“Midas：使用近似记忆系统的模型反演防御”，载于2020年亚洲硬件安全与可信研讨会（AsianHOST）。IEEE，2020年，第1–4页。[25] C. Dwork，“差分隐私”，载于自动机、语言与编程国际研讨会。Springer，2006年，第1–12页。[26] Y. Zhang、R. Jia、H. Pei、W. Wang、B. Li 和 D. Song，“秘密揭示者：针对深度神经网络的生成模型反演攻击”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第253–261页。[27] Y. Zhang、Z. Lv、H. Wu、S. Zhang、P. Hu、Z. Wu、H.-y. Lee 和 H. Meng，“Mfa-conformer：用于自动说话人验证的多尺度特征聚合 Conformer”，arXiv预印本 arXiv:2203.15249，2022年。[28] B. Desplanques、J. Thienpondt 和 K. Demuynck，“Ecapa-tdnn：基于时延神经网络的说话人验证中的通道注意力强化、传播与聚合”，arXiv预印本 arXiv:2005.07143，2020年。