声音事件嵌入揭示了什么？审计音频表征中的说话人信息泄漏

Bhawana Chhaglani

宾夕法尼亚州立大学
bbc5655@psu.edu

arXiv:2611.31043v1 [cs.SD] 23 Nov 2026

摘要

预训练音频嵌入日益被用作声音事件检测（SED）及其他常开式声学感知任务的中间表示。尽管选择这些表征是为了提升下游任务的效用，但它们也可能保留与预期用途无关的敏感信息，即使原始音频未被暴露，也会带来隐私风险。本研究提出了一个实证框架，采用两种互补方法审计预训练音频嵌入中的隐私泄漏。首先，通过测量音频嵌入与专门用于编码敏感信息的表征之间的统计依赖，量化表示级泄漏。其次，通过训练轻量级对抗探测器直接从嵌入中恢复敏感属性，测量攻击级泄漏。我们将此框架应用于四种广泛使用的音频表征（CLAP、PANNs、AST 和 YAMNet），并评估说话人身份和性别信息的泄漏情况。结果发现，说话人识别的平衡准确率介于 0.43 至 0.72 之间，性别信息也能被高度准确地恢复。此外，表示级和基于攻击的测量结果呈现一致趋势：CLAP 泄漏的说话人身份信息最多，而 YAMNet 最少。这些结果表明，中间音频表示并非天然具备隐私保护能力的接口，并说明在为声学感知系统选择表征时，除了下游效用之外，还应将隐私泄漏作为一项明确的评估标准。

引言

音频感知越来越多地应用于智能音箱、可穿戴设备及其他始终开启的设备，用于识别警报声、玻璃破碎声、咳嗽声等环境声音和其他声学事件（Amazon Staff 2024；Heittola 等人 2013；Chhaglani 等人 2024）。现代声音事件检测（声音事件检测）系统通常不会直接处理原始音频，而是依赖预训练音频编码器将输入信号 X 转换为中间表示 Z，随后将其发送至云端，用于下游任务 T，如图 1 所示。CLAP（Elizalde 等人 2023）、PANNs（Kong 等人 2020）、AST（Gong、Chung 和 Glass2021）和 YAMNet（Korkmaz 2025）等模型提供了通用表示，使得构建此类系统无需从头训练音频编码器。这些系统已被证明在各种应用中效果显著。例如，YAMNet 已用于心血管疾病分类（Barnawi 2023）、语音情感识别（Boateng 和 Kowatsch 2020）、COVID 检测（Sobahi 等人 2022）以及帕金森病语音识别（Syed、Memon 和 Memon 2020）。类似地，CLAP 已用于心率估计和听诊（Nie 等人 2025），以及声音事件检测（Elizalde 等人 2023）（包括紧急声音）。

然而，尽管这些嵌入旨在捕获对声学理解有用的信息，但它们是根据完整的声学信号计算得到的，也可能编码与预期任务无关的信息。当环境中出现语音时，这可能包括有关被录音人员的敏感信息，即使原始音频本身从未被公开（Williams 等（2023年）；Chhaglani、Gummeson 和 Shenoy（2024年）；Chen、Adcock 和 Krishnagiri（2008年）；Boovaraghavan 等（2024））。这引出了隐私感知音频传感中的一个根本问题：我们如何知道音频嵌入是否泄露了私人信息？

理想情况下，一种表示应保留与其预期任务相关的信息 T，同时舍弃与敏感属性相关的信息 S，即 $I(Z; S) \ll I(Z; T)$。然而，音频嵌入通常是根据下游效用来选择的，而对于其中仍编码了哪些非预期信息，人们的关注要少得多。因此，两个功能相近的嵌入可能会暴露出关于用户的截然不同数量的信息。这对于中间表示尤为重要，因为人们往往认为它们比原始音频更安全，却将其存储、传输或提供给下游应用访问。

在这项工作中，我们通过审查常用音频嵌入中的说话人隐私来探究这一问题。我们使用两种互补的泄漏度量，研究了四种常用的预训练音频表示：CLAP、PANNs、AST 和 YAMNet。首先，我们提出一种表示层面的度量方法，用于衡量音频嵌入中的信息与专门用于编码说话人身份的表示有多大重叠。我们提取

同一音频中的 ECAPA-TDNN（Guo 等，2024）说话人嵌入，并测量

版权所有 © 2027年，人工智能促进协会（www.aaai.org）。保留所有权利。

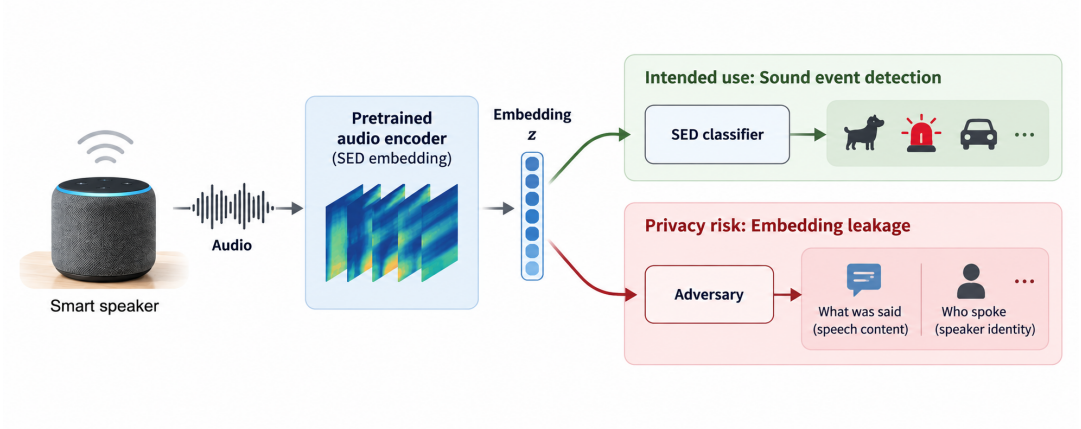

图 1：用于智能音箱声音事件检测的音频嵌入

这两个表征空间之间的统计依赖，使用典型相关分析（CCA）和互信息（MI）估计。其次，我们进行攻击层面的评估：攻击者仅接收音频嵌入，并训练轻量级分类器来推断敏感属性。我们考察说话人识别和性别推断，以此衡量嵌入中保留的信息是否能被实际利用。

我们使用涵盖五位说话人（三位男性和两位女性）的 Common Voice 数据集评估这些方法。结果表明，不同音频嵌入之间存在显著且一致的差异。对于说话人识别，每位说话人仅使用 2 秒注册音频时，平衡准确率从 YAMNet 的 0.43 到 CLAP 的 0.72 不等。我们的表征层面分析呈现出相同的整体趋势：CLAP 对说话人专属表征的依赖程度最高，而 YAMNet 最低。在所评估的表征中，也都能恢复出性别信息，其中 CLAP 的信息泄漏最严重。

我们的研究结果揭示了隐私敏感型感知系统在选择音频表征时一个被忽视的维度。不能仅仅因为丢弃了原始音频，就认为中间嵌入具备隐私保护能力：针对声学效用优化的表征，仍可能保留信号中人物的大量信息。更广泛地说，我们的结果表明，应从两个维度评估音频嵌入：对预期任务的效用 和敏感信息泄漏 ，并将隐私作为明确的模型选择标准。此类审查有助于开发者识别出既能保留应用所需功能、又能最大限度减少不必要用户信息暴露的表征。

背景

音频感知中的隐私可以从处理流程的不同层面进行考察：信号级、特征级和模型级隐私。信号级隐私。 信号级方法旨在直接抑制音频信号中的敏感信息，同时保留对声学感知有用的信息。这包括低通滤波等技术（Pohlhausen 和 Bitzer（2025）；Iravantchi 等（2021年）、子采样（Mollyn 等（2022年））、移除可能包含语音的片段（Boovaraghavan 等（2024年））等。例如，Kirigami（Boovaraghavan 等（2024年））研究了信号级隐私技术中的语音内容泄露问题。特征级隐私。特征级方法会仔细提取与任务相关的音频信号特征。例如，Symdetector（Sun 等（2015））从音频中提取手工设计特征，以检测咳嗽和打喷嚏。FeatureSense（Chhaglani 等（2025年））评估隐私（说话人）泄露，并研究低层手工设计声学特征中的隐私与效用权衡，结果表明，用于声音分类的特征也可能保留敏感信息。这类分析有助于找出能够平衡下游感知效用与隐私的特征表示。模型级隐私。模型级隐私技术旨在保障机器学习模型的隐私，防止模型学习或暴露敏感信息，同时保留其在预期任务中的效用。常见方法包括差分隐私（Paul，2025年）、对抗学习和隐私保护表征学习（Gharib 等（2024年））。与修改模型训练的方法不同，我们的工作审查现有预训练音频编码器生成的中间嵌入所涉及的隐私问题。具体而言，我们评估这些表示是否保留了可被攻击者恢复的说话人身份和性别信息。我们的工作将隐私审计从原始音频和手工设计特征扩展到了中间神经表征。

隐私评估中的挑战。隐私泄露的评估仍然具有挑战性，因为即使敏感信息无法被直接观察到，也可能保留在音频表征中。因此，仅依靠主观听测是不够的，因为人类无法理解的信息仍可能被机器学习模型恢复（Pohlhausen 和 Bitzer（2025年）；Boovaraghavan 等（2024年））。隐私风险还取决于攻击者的能力、可用的训练数据，以及所针对的敏感属性（如语音、说话人或上下文）。此外，统计依赖并不一定表明相关信息在实践中是否能够被利用。这些挑战促使我们采用互补的评估方法，同时衡量表征中保留的信息及其通过对抗性攻击的可恢复性。因此，我们的威胁模型假设存在一名

能够访问预训练音频嵌入、但无法获取原始音频的攻击者，并使用机器学习探针推断说话人身份及其属性。

音频嵌入中的隐私泄漏审计

问题表述

考虑一个音频编码器 f，它将输入音频信号X 映射为中间表示 $Z = f(X)$。该表示会提供给执行预期任务 T的下游应用，例如声音事件检测。然而， Z 也可能同时保留与敏感属性 S 有关的信息，而这些信息对于预期任务并非必需，例如说话人身份或性别。从概念上讲，保护隐私的表示应最大限度地保留与预期任务相关的信息，同时最大限度地减少关于敏感属性的信息： $I(Z;\ S) \ll I(Z;\ T)$。因此，当敏感信息仍编码在 Z 中，且尽管与预期应用无关，仍可被检测或恢复时，我们认为该嵌入存在隐私泄漏。我们从两个互补角度评估此类泄漏：表示级依赖性和攻击级可恢复性。

表示级泄漏

我们的第一种方法用于衡量音频表示是否包含与某种敏感属性相关的信息，而不显式训练攻击者来预测该属性。我们使用预训练 ECAPA-TDNN 说话人编码器，获得参考表示 Z_S，该表示经过专门优化以实现说话人判别。给定同一条语音中的音频嵌入 Z和对应的说话人嵌入 Z_S，我们衡量这两个表示空间之间的依赖性。

由于这些表征的维度不同，我们首先对其进行标准化并应用主成分分析（PCA），然后进行典型相关分析（CCA），以识别相关性最高的成分。接着，我们使用 k 近邻估计器来估算成对典型成分之间的互信息。依赖性越高，表明 Z 保留了更多与 Z_S 对齐的信息。我们将其用作表征层面的说话人信息泄漏指标，而非对 I（Z;S）的直接估计。

攻击级泄漏

我们的第二种方法探讨编码在 Z 中的敏感信息是否会被攻击者利用。我们假设攻击者可以访问中间嵌入，但无法获取原始音频，并训练一个轻量级分类器 g： $Z \to S$ 来恢复某个敏感属性。从信息论视角来看，敏感信息泄漏可以表示为 $I(Z;S) = H(S) - H(S|Z)$：观察 Z使 S的不确定性降低得越多，该表示揭示的敏感信息就越多。我们的攻击层面评估通过测量攻击者从 Z中推断 S 的准确程度，将这一概念付诸实践。

我们评估两种敏感属性。对于说话人身份，我们使用固定数量的登记音频训练攻击者，并测试其能否识别未见过语句的说话人。对于性别，我们基于冻结表示训练一个二分类器。我们报告平衡准确率以考虑类别不平衡，并在多个随机划分上重复每项实验。

评估

实验设置

数据集。 我们使用 Mozilla Common Voice 中的 100 段两秒语音片段评估隐私泄漏，这些片段来自五位说话人（三位男性、两位女性）。为确保比较条件受控，所有嵌入均使用相同的音频片段。

音频嵌入。 我们评估四种广泛使用的预训练音频表示：CLAP、PANNs、AST 和 YAMNet。对于每段片段，我们提取相应的冻结表征，不对编码器进行微调。对于输出帧级表征的模型，我们应用时间均值池化，以获得固定维度的嵌入。最终嵌入维度分别为：CLAP 为 512，PANNs 为 2048，AST 为 768，YAMNet 为 1024。

敏感表示与攻击。 对于表示级泄漏，我们使用预训练的 ECAPA-TDNN 说话人识别模型，从每个片段中提取说话人专用嵌入，并衡量其与每个音频嵌入的依赖关系。对于攻击级泄漏，我们在冻结的嵌入上训练逻辑回归探针，以预测说话人身份和性别。我们报告平衡准确率，并在多个随机划分上重复实验。对于说话人识别，我们还会改变攻击者可获得的注册音频量，以刻画说话人信息被利用的速度。

实现细节。 对于表示级泄漏，我们使用 StandardScaler 分别对音频嵌入和 ECAPA-TDNN 嵌入进行标准化。随后，我们对每种表示应用主成分分析，各保留 10 个成分，然后使用 5 个典型成分和最多 2,000 次迭代进行典型相关分析。我们使用 mutual_info_regression ($k = 3$) 中实现的 k近邻估计器，估计对应典型成分之间的互信息。我们报告 5 个成分的平均互信息。统计显著性通过对配对的说话人嵌入进行 100 次随机置换来评估。

攻击配置。我们使用标准化的冻结音频嵌入训练逻辑回归分类器，并采用平衡类别权重、L2正则化、$C = 1.0$，迭代次数最多为 3,000 次。对于说话人识别，我们随机为每位说话人保留 5 段音频片段用于测试，并额外选择 1 段作为注册样本。我们使用随机种子 42–51 重复此过程 10 次，并报告平衡准确率的均值和标准差。

对于性别推断，我们采用分层随机划分，将 75% 的音频片段用于训练，25% 用于测试。使用随机种子 42，基于 20 次随机划分重复实验，并报告平均平衡准确率和标准差。这些划分以音频片段为单位，因此训练集和测试集中可能包含相同的

表 1：通过与 ECAPA-TDNN 说话人嵌入的依赖性测量表征层面的说话人信息泄漏。数值越高，表示对说话人专用表征的依赖性越强。

嵌入	平均典型相关分析	平均互信息	互信息总和
CLAP	0.733	0.448	2.242
PANNs	0.671	0.325	1.624
AST	0.647	0.276	1.379
YAMNet	0.646	0.264	1.319

说话人。因此，性别结果衡量的是所评估说话人的属性可恢复性，而非对未见说话人的泛化能力。

音频嵌入中的隐私泄漏：结果

表示层面的说话人信息泄漏。我们首先考察通用音频嵌入是否保留了与说话人身份相关的信息。表 1 报告了各类音频表示与针对说话人优化的 ECAPA-TDNN 嵌入之间的依赖性。我们观察到不同表示之间存在明显差异。CLAP 的依赖性最高，平均 CCA 为 0.733，平均 MI 为 0.448；按平均 MI 排序，其后依次为 PANNs（0.325）、AST（0.276）和 YAM-Net（0.264）。这些结果表明，为通用声学理解而设计的表示可能保留大量说话人相关信息，且保留的信息量因嵌入模型而异。

说话人识别攻击。接下来，我们测试表示层面观察到的说话人信息能否被攻击者利用。图 2 展示了攻击者对每位说话人仅获得一个 2 秒注册片段时的说话人识别效果。在五分类识别中，四种表示的表现均超过 0.20 的随机猜测水平。CLAP 支持的攻击效果最强，仅凭 2 秒注册音频便达到 0.72 的平衡准确率；其后依次为 PANNs（0.52）、AST（0.51）和 YAMNet（0.43）。表示层面和攻击层面的测量结果总体趋势一致：CLAP 的说话人信息泄漏最严重，YAMNet 最轻微。不过，与观测到的变异相比，中间几种模型之间的差异较小，因此仍需进行更大规模的评估，才能确定表示级依赖性是否能够可靠地预测攻击表现。

性别推断。说话人身份并非这些表示保留的唯一敏感信息。图3展示了如何从相同的冻结嵌入中推断性别。CLAP在我们的评估中实现了完美分类（平衡准确率为1.00），而AST、PANNs和YAMNet的平衡准确率分别为0.88、0.83和0.73，均高于0.50的二分类随机基准。结合说话人识别结果，这些发现表明，旨在用于通用声学任务的音频嵌入可能会暴露音频中个体的多种属性，并且不同表示之间的泄漏程度差异显著。这些差异表明，仅凭嵌入维度无法解释隐私泄漏。预训练目标、架构，

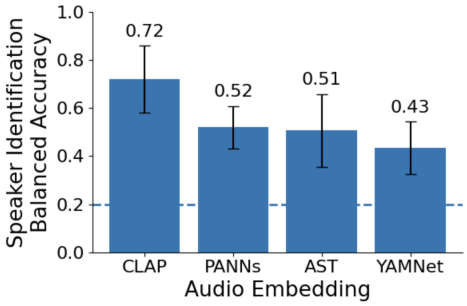

图2：针对音频嵌入的说话人识别攻击。误差条表示多次运行的标准差；虚线表示随机猜测性能（0.20）。

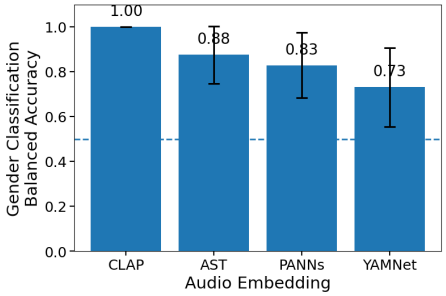

图3：基于冻结音频嵌入的性别推断。误差条表示多次运行的标准差；虚线表示随机猜测性能（0.50）。

以及表示提取策略都可能导致观察到的差异，尽管要分离它们各自的影响仍需进行受控实验。

结论与未来工作

我们提出了一个经验性框架，利用互补的表征层面和攻击层面测量，对预训练音频嵌入中的隐私泄漏进行审计。我们的评估表明，用于通用声学理解的表征可能保留大量与说话人相关的信息，即使无法获取原始音频，也能够推断说话人身份和性别。重要的是，表征层面分析和说话人识别分析在所评估的嵌入中呈现出一致的趋势：在我们的实验中，CLAP表现出最高的说话人信息泄漏，而YAMNet最低。这些发现表明，不应假定中间音频表征具备隐私保护能力，并促使我们在为音频感知系统选择嵌入时，将隐私泄漏作为一项明确标准。我们的研究是针对少量说话人和表征开展的初步审计。未来的工作将评估规模更大、更加多样化的说话人群体，研究其他敏感属性（例如语音内容），并采用更多音频编码器。我们还计划联合刻画下游SED效用和敏感信息泄漏，从而分析音频表征的隐私-效用权衡。

参考文献

Amazon Staff。2024年。关于Alexa Emergency Assist（这项亚马逊服务可保障您和家人的安全）的一切须知。About Amazon。访问日期：2026年11月19日。

Barnawi, S. A. A. 2023年。预训练CNN模型用于心血管疾病分类的评估：一项基准研究。*Information SciencesLetters*，12(7)：3317–38。Boateng, G.；以及Kowatsch, T. 2020年。使用多模态融合和迁移学习对老年人进行语音情感识别。载于*2020*年国际多模态交互会议论文集，12–16。Boovaraghavan, S.；Zhou, H.；Goel, M.；以及Agarwal, Y. 2024年。Kirigami：利用音频进行隐私保护的轻量级语音滤波活动识别。*ACM*交互式、移动、可穿戴与普适技术论文集，8(1)：1–28。Chen, F.；Adcock, J.；以及Krishnagiri, S. 2008年。音频隐私：在保留环境声音的同时降低语音可懂度。载于第*16*届*ACM*国际多媒体会议论文集，733–736。Chhaglani, B.；Gomasta, S. S.；Agarwal, Y.；Gummeson, J.；以及Shenoy, P. 2025年。FeatureSense：保护持续开启音频感知系统中的说话人属性。*arXiv*预印本 *arXiv:2505.24115*。Chhaglani, B.；Gummeson, J.；以及Shenoy, P. 2024年。迈向隐私保护音频分类系统。*arXiv*预印本 *arXiv:2404.18002*。Chhaglani, B.；Nemati, E.；Chandrashekhara, S.；Kuang, J.；以及Gao, A. 2024年。Cough-DL：用于耳戴式咳嗽检测的深度学习模型。载于*2024*年第*46*届*IEEE*医学与生物学工程学会国际年会（*EMBC*），1–5。IEEE。Elizalde, B.；Deshmukh, S.；Al Ismail, M.；以及Wang, H. 2023年。通过自然语言监督学习音频概念的CLAP。载于*ICASSP 2023—2023*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），1–5。IEEE。Gharib, S.；Tran, M.；Luong, D.；Drossos, K.；以及Virtanen, T. 2024年。面向音频稳健隐私保护的对抗表示学习。*IEEE*开放信号处理期刊，5：294–302。Gong, Y.；Chung, Y.-A.；以及Glass, J. 2021年。AST：音频频谱图变换器。*arXiv*预印本 *arXiv:2104.01778*。Guo, J.；Zhu, J.；Lin, S.；以及Shi, F. 2024年。用于说话人识别的ECAPA-TDNN嵌入。载于*2024*年第*5*届人工智能、网络与信息技术国际研讨会（*AINIT*），1488–1491。IEEE。Heittola, T.；Mesaros, A.；Eronen, A.；以及Virtanen, T. 2013年。上下文相关声音事件检测。*EURASIP*音频、语音与音乐处理期刊，2013(1)：1。Iravantchi, Y.；Ahuja, K.；Goel, M.；Harrison, C.；以及Sample, A. 2021年。Privacymic：利用不可听频率实现隐私保护的日常活动识别。载于*2021*年*CHI*计算机系统中的人为因素会议论文集，1–13。

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumb- ley, M. D. 2020年。Panns：用于音频模式识别的大规模预训练音频神经网络。*IEEE/ACM* 音频、语音与语言处理汇刊，28：2880–2894。Korkmaz, Y. 2025年。基于音频信号的情绪识别中 YAMNet 和 VGGish 网络的性能分析。*European Journal of Technique (EJT)*，15(2)：251–260。Mollyn, V.; Ahuja, K.; Verma, D.; Harrison, C.; and Goel, M. 2022年。SAMoSA：利用运动和子采样音频感知活动。*ACM* 交互式、移动、可穿戴与普适技术会议论文集，6(3)：1–19。Nie, J.; Tran, D. T.; Thakkar, K.; Kowtha, V.; Huang, J.; Avendano, C.; Azemi, E.; and Mitra, V. 2025年。用于通过听诊估计心率的基础模型隐藏表示。*arXiv* 预印本 *arXiv:2505.20745*。Paul, A. 2025年。安全与隐私保护的音频处理技术。Pohlhausen, J.; and Bitzer, J. 2025年。重新审视低频语音信号的隐私：探索重采样方法、评估场景和说话人特征。*arXiv* 预印本 *arXiv:2508.02483*。Sobahi, N.; Atila, O.; Deniz, E.; Sengur, A.; and Acharya, U. R. 2022年。利用分形维数和结合 Grad-CAM 的视觉Transformer对咳嗽声音进行可解释的 COVID-19 检测。生物控制论与生物医学工程，42(3)：1066–1080。Sun, X.; Lu, Z.; Hu, W.; and Cao, G. 2015年。SymDetector：使用智能手机检测与声音相关的呼吸道症状。在*2015* 年 *ACM* 普适与泛在计算国际联合会议论文集，97–108。Syed, Z. S.; Memon, S. A.; and Memon, A. L. 2020年。用于识别帕金森病语音的深度声学嵌入。国际高级计算机科学与应用期刊，11(10)：726–734。Williams, J.; Pizzi, K.; Das, S.; and Noé, P.-G. 2023年。语音与音频内容隐私面临的新挑战。*arXiv* 预印本 *arXiv:2301.08925*。