面向多语言说话人嵌入的因果解耦对比学习

Mariëtte Olijslager[1], Seyed Sahand Mohammadi Ziabari[1,2], Ali Mohammed Mansoor Alsahag[1]1信息学研究所，阿姆斯特丹大学，阿姆斯特丹，荷兰2计算机科学与技术系，纽约州立大学帝国州立大学，萨拉托加斯普林斯，纽约州，美国
mariette.olijslager@student.uva.nl,sahand.ziabari@sunyempire.edu,a.m.m.alsahag@uva.nl

arXiv:2602.01363v1 [cs.SD] 1 Feb 2026

摘要
自监督说话人嵌入被广泛应用于说话人验证系统，但已有研究表明，它们往往编码敏感的人口统计属性，从而引发公平性与隐私方面的担忧。本文研究经SimCLR训练的说话人嵌入中人口统计信息——具体而言，性别、年龄和口音——的存在程度，以及能否在不严重降低说话人验证性能的前提下缓解此类泄漏。我们研究两种去偏策略：通过梯度反转进行的对抗训练，以及显式分离人口统计信息与残差信息的因果瓶颈架构。人口统计信息泄漏使用线性和非线性探测分类器进行量化，而说话人验证性能则使用ROC-AUC和等错误率进行评估。我们的结果表明，性别信息在基线嵌入中被强烈且线性地编码，而年龄和口音则较弱，且主要以非线性方式表示。对抗去偏能够减少性别信息泄漏，但对年龄和口音的影响有限，并与验证准确率之间形成明显权衡。因果瓶颈进一步抑制了人口统计信息，尤其是在残差表示中，但会造成显著的性能下降。这些发现凸显了在自监督说话人嵌入中缓解人口统计信息泄漏的根本局限，并阐明了当前去偏方法所固有的权衡。

关键词
说话人验证、自监督学习、对比学习、人口统计信息泄漏、对抗去偏

1 引言

近年来，深度学习技术显著推动了说话人识别与验证领域的发展。如今，基于大量语音语料库训练的大规模模型能够捕获高度可区分的声学表示，即使在具有挑战性的声学条件下，也能实现出色的说话人识别 [29, 30]。然而，随着这些模型日益强大，人们开始担忧模型会在潜在表征中无意编码性别、年龄或口音等敏感的人口统计信息。这种现象称为人口统计信息泄漏，可能危及公平性、隐私和伦理合规性。它可能导致不同人口统计群体之间出现系统性偏差，或使人们能够从表面上已匿名化的嵌入中推断受保护属性 [32, 33]。因此，解决此类泄漏对于确保公平且负责任地部署语音技术至关重要 [34]。重要的是，此类泄漏不仅可能源于直接的统计相关性，也可能源于将人口统计属性与学习到的表征联系起来的潜在因果机制，因此有必要探索超越基于相关性的缓解方法。

历史上，说话人嵌入通常通过有监督分类目标进行学习，这类目标需要大量带标签的说话人数据，并可能无意中编码与非说话人属性之间的虚假相关性。为缓解这些局限，近期研究提出了 SimCLR 和 wav2vec 等自监督对比学习方法。这些方法表明，无需说话人标签，通过最大化同一输入的不同增强视图之间的一致性，即可学习到有意义的语音表征 [31]。尽管对比学习方法具有良好的可扩展性和强大的表征能力，但它们本身并不能确保对敏感人口统计属性保持不变。因此，即使下游说话人验证并不需要人口统计信息，这些信息仍可能保留在学习到的嵌入中，从而可能损害公平性和隐私 [51]。这促使我们系统地研究人口统计信息泄漏在对比式说话人嵌入中的表现形式，以及如何加以缓解。

尽管越来越多的证据表明说话人表征中存在人口统计信息泄漏，但系统量化或缓解自监督对比学习嵌入中这种偏差的研究仍相对较少。以往大多数工作都聚焦于监督式说话人识别中的对抗式或多任务去偏策略，揭示了固有的公平性与效用权衡，以及这些策略在彻底抑制人口统计信息方面能力有限 [40, 41]。此外，研究者也提出了旨在减少人口统计差异的架构调整，并将其用于监督式场景 [42]。然而，这些方法通常无法直接用于自监督对比学习表征，因为后者的学习目标和归纳偏置在根本上有所不同。在对比学习中，学习目标鼓励表征在不同话语之间具有尽可能强的区分能力，从而隐式强化数据中存在的任何人口统计规律，并形成纠缠嵌入，使任务相关属性与敏感属性共同编码其中。因此，人口统计信息泄漏成为一个表征学习问题，而非下游策略问题。尤其是，在自监督说话人嵌入模型中，利用显式结构化或因果驱动的瓶颈架构来分离和控制人口统计信息，仍鲜有探索。

本文探讨了一种比较框架，用于缓解自监督说话者嵌入中的人口统计信息泄漏。首先，

采用对比学习方法 SimCLR，训练模型以学习说话人判别性表征，作为基线。其次，应用对抗去偏来抑制人口统计信息，并使用线性和非线性探测分类器评估泄漏，以衡量超出线性可分性的人口统计信息泄漏。第三，在表征流水线末端引入因果瓶颈层，明确约束敏感属性流向最终嵌入的信息流。通过将人口统计属性建模为因果上游因素，并限制它们对所学表征的影响，因果瓶颈为仅使用对抗目标来缓解人口统计信息泄漏提供了一种新颖的替代方案。

本文研究自监督说话人嵌入中的人口统计信息泄漏，并评估能否通过对抗去偏和因果瓶颈架构缓解此类泄漏，同时保持说话人验证性能。研究首先分析经 SimCLR 训练的说话人嵌入如何编码与性别、年龄和口音相关的人口统计属性，并使用线性和非线性探测分类器量化泄漏的程度和结构。随后，研究考察对抗去偏在抑制人口统计信息方面的有效性，并评估这一干预对线性编码和非线性编码属性以及说话人验证准确率的影响。最后，本文评估引入因果瓶颈层能否在单独使用对抗去偏的基础上进一步抑制人口统计信息泄漏，尤其是非线性人口统计信息泄漏，并分析由此带来的说话人验证性能权衡。

通过综合评估说话人验证准确率、人口统计探测和公平性相关指标，本文旨在刻画不同去偏策略在自监督说话人表征中的有效性与局限性。通过对比对抗目标与因果驱动的架构约束，本研究有助于深入理解人口统计信息泄漏如何产生，以及如何在现代对比学习框架中加以缓解。

2 相关工作

本节回顾与语音表征中的人口统计信息泄漏及现有缓解策略相关的先前工作。我们首先总结相关证据，说明现代说话人嵌入模型会编码性别、年龄和口音等敏感人口统计属性。随后，我们回顾语音领域中的对比学习方法和公平性感知表征学习，并进一步介绍将受保护属性与学习到的嵌入解耦的因果方法和对抗方法。这一脉络为本研究的方法选择提供了依据，并指出了本研究旨在解决的不足。

2.1 说话人嵌入中的人口统计信息

理解说话者嵌入如何编码人口统计信息，对于设计公平且注重隐私的说话者验证系统至关重要。近期研究考察了流行嵌入模型在多大程度上保留性别、年龄和口音线索，从而凸显潜在的人口统计信息泄漏。无论是WavLM [44]（一种自监督语音模型）还是ECAPA-TDNN [30]（一种有监督的说话者验证编码器），都会生成保留人口统计线索的稠密嵌入，但这两种架构都并非为抑制此类信息而设计，因此在缓解人口统计信息泄漏方面并不被视为最先进水平。先前的分析已量化不同人口统计属性在这些嵌入中的编码方式。性别几乎线性可分：简单的线性探针准确率超过99.8%，表明嵌入空间中单一主导方向承载了性别信号。年龄的编码更为分散，回归探针的平均绝对误差约为4.9年，说明年龄信息分布在多个维度上，需要非线性映射（例如多层感知机或岭回归）才能有效恢复 [44]。口音介于这两个极端之间：CommonAccent等分类基准表现出较强性能，但始终低于性别预测，表明口音线索存在，但分散在多个维度上。

这些发现呈现出一致的规律：性别最容易从说话人嵌入中预测，年龄最难，口音则介于两者之间。由于这些编码器具有深层且非线性的结构，其潜在表征本身更为复杂；性别信息可以通过线性分类器读出，而年龄和口音通常需要非线性探测。重要的是，这种非线性信息泄漏促使我们在下游实验中同时评估线性与非线性探针，以全面量化嵌入中的人口统计信息。

总体而言，这些分析表明，原始 WavLM 或 ECAPA-TDNN 嵌入在人口统计特征方面存在纠缠。因此，要实现真正的隐私保护或公平性，需要引入额外目标，例如对抗去偏、互信息最小化（如 WavShape [43]），或通过对比式设计来强制实现不变性。这促使本研究探索对比式自监督说话人嵌入，并引入因果瓶颈架构，旨在明确地将说话人判别信息与人口统计因素分离。

2.2 说话人嵌入的对比学习方法

FairASR [56] 表明，对抗性人口统计信息抑制可以缩小自动语音识别（ASR）输出中的人口统计差距，同时使总体词错误率（WER）与基线模型相当。不过，该方法依赖预训练期间使用的人口统计标签，只报告总体差异缩减，未提供详细的分组统计数据，且评估主要基于域内语料库。后续研究指出，自动语音识别的公平性取决于具体情境，在某一人口统计维度上的改进未必能推广到不同数据集或分布偏移的情形 [57]。近期的一种对比学习方法将性别替换话语配对，以生成性别不变嵌入；该方法在跨域公平性方面表现出色，同时只需牺牲极少的词错误率 [58]。

尽管取得了这些进展，先前方法的适用范围仍然有限：它们通常只针对单一人口统计属性，关注自动语音识别而非说话人验证，也未系统评估非线性人口统计信息泄漏或交叉性效应。

相比之下，本文研究自监督说话人嵌入（例如基于 SimCLR 的嵌入），并将对抗去偏

与因果瓶颈相结合，以减轻多个人口统计维度（性别、年龄、口音）及其交叉处的信息泄漏。我们使用线性与非线性探针量化泄漏，并通过自助法置信区间提供严谨的不确定性估计。与 FairASR 及其他以自动语音识别为重点的方法不同，我们的工作专门将对抗去偏和因果瓶颈应用于说话人嵌入，并系统地研究非线性泄漏和交叉性效应。

2.3 因果解缠

为了进一步抑制学习到的表征中不需要的人口统计信息，可以通过因果瓶颈层纳入对抗去偏技术。其基本思想在领域对抗神经网络（DANN）[36],中提出，将主编码器与一个辅助对抗器耦合，该对抗器被训练为根据潜在特征预测领域或受保护属性，同时梯度反转层迫使编码器去除此类信息。

Elazar 和 Goldberg [37] 的后续工作将这一概念扩展到文本数据，表明对抗目标能够减少但并不能完全消除人口统计信息泄漏，从而凸显了使用独立的攻击者分类器验证残余偏差的重要性。近期研究强调了用于偏差缓解的因果方法与表征层面方法。Koçadağ 等人 [52] 提出了一种用于词嵌入中交叉性偏见缓解的基于结构因果模型的框架，表明显式建模因果关系可提升连贯性与公平性。类似地，Zhu 等人 [53] 利用结构因果模型引入了任务自适应去偏，展示了任务性能与偏差缓解之间固有的权衡。

除因果建模外，Garbat 等人 [54]研究了沿社会根植维度对上下文嵌入进行去偏，而 Naranbat 等人 [55] 表明，公平性结果对评估设计和领域选择极为敏感。

2.4 通过对抗去偏和瓶颈对表征进行去偏

Edwards 和 Storkey [38] 以及 Zhang 等人 [39]的早期研究提出了使用梯度反转层的对抗目标，以抑制受保护属性，同时保留任务相关信息。这些方法通常需要人口统计标签，并会在公平性与预测性能之间引入权衡。

在语音领域，FairASR [56] 将多人口统计维度对抗目标与用于自动语音识别的对比学习相结合，以减少人口统计差异，但依赖域内评估和汇总指标。通过性别互换配对进行对比学习 [58] 则提供了另一种方案，无需明确的人口统计标签即可提升公平性。

作为对抗方法的补充，FairIB [59] 等信息瓶颈方法会压缩嵌入，在保留任务相关信息的同时减少人口统计信息泄漏。然而，这些方法大多在语音领域之外进行探索，或未与对比目标结合。

我们的工作在一个统一框架中整合了对抗去偏与因果瓶颈，用于自监督说话人嵌入，从而能够详细分析线性和非线性人口统计信息泄漏及其与说话人验证性能之间的权衡。

3 方法

本节介绍用于研究自监督说话人嵌入中人口统计信息泄漏的实验设置。我们首先介绍数据集和预处理流程，然后详细说明特征提取和训练流程，最后介绍说话人验证和人口统计探测的评估方案。

3.1 数据集

3.1.1 数据集属性。本文使用 Mozilla Common Voice 数据集，因为该数据集提供已验证语音片段及性别、年龄和口音等元数据，这些正是本研究旨在考察的人口统计特征 [50]。可从 Mozilla Data Collective 网站下载该数据集的 *tar.gz* 文件 [50] ，下载时需使用 API密钥。具体而言，我们使用了 ***Common Voice Scripted Speech 23.0*** – 英语 数据集，其中包含来自 98,579 位说话者的 2,541,751 个片段（共 3,746 小时录制语音，其中 2,682 小时经过验证）。我们仅使用通过 validated.tsv 文件标识的已验证片段，以确保标签可靠性。

3.1.2 预处理。39%的说话者性别未定义、属于非二元性别或跨性别。我们将这些说话者排除在外，因为纳入他们可能会混淆线性探针：嵌入可能反映多种性别信号或噪声，而非明确的二元性别差异。因此，我们聚焦于二元性别类别（男性和女性），从而更可靠地测量性别信息泄露。36%的说话者缺少年龄值。为确保基于准确的元数据进行年龄相关探测，我们将这些说话者移除。原始的十年年龄区间被划分为三类：青年、成年人和老年人。这种分组考虑了感知局限：人类听者往往无法区分细微的年龄差异，例如三十多岁和四十多岁之间的差别，同时又保留了有意义的年龄变化。对于口音，我们将自由文本标签规范为五类：美国、英格兰、加拿大、澳大利亚/新西兰、印度/东南亚。这样可以减少标注不一致带来的噪声，并确保每组有足够的样本以进行可靠评估。我们排除这些类别以外的口音。少于两条话语的说话者也会被移除，因为说话人验证需要比较每位说话者至少两条不同话语的嵌入。

3.1.3 说话人分布与数据划分。预处理后，数据集包含 11,209 名独立说话者，每人至少有两条话语。最后，数据集划分为训练集、验证集和测试集，分别包含 8,967、1,121 和 1,121 名说话者。人口统计属性存在类别不平衡（见表 1），这反映了 Common Voice 数据集中说话者的自然分布。为避免人为改变数据分布，本研究未采用重采样，而是在训练过程中使用了不平衡感知损失。因此，这些结果被视为人口统计信息泄漏的上界，反映了从冻结的

表 1：处理后的数据集中说话者的人口统计分布。百分比表示各类别中说话者所占的比例。

类别	说话者	百分比 (%)
性别		
男性 男性化	8968	80.01
女性 女性化	2241	19.99
年龄组		
青年	6227	55.52
成年人	4388	39.12
老年人	601	5.36
口音组		
USA	5737	51.18
英格兰	1687	15.05
印度／南亚	1661	14.82
其他（不包括）	835	7.45
加拿大	655	5.84
澳大利亚／新西兰	635	5.66

说话者嵌入中使用简单分类器（例如线性探针）推断敏感属性的最大程度，而非所能达到的最佳分类性能。

3.2 音频处理

首先，我们将采样率设为 16 kHz，以捕获与人类语音最相关的频率范围，同时丢弃对说话人身份贡献甚微、却会增加计算成本的高频成分。这确保了各项实验的输入分辨率一致，并使模型与说话人表征学习中常用的语音频谱特征相契合 [29, 47]。我们使用 25 ms 的窗口和 10 ms 的帧移提取频谱特征。这一配置兼顾了频率分辨率和时间分辨率：窗口长度能够捕捉稳定的声道特征，而帧移则保留短时的时间变化。这样的平衡对对比学习十分重要，因为表征应对局部时间变化保持不变，同时保留说话人特有的线索 [29, 48]。我们采用 64 个梅尔滤波器组，在频谱表达能力与模型容量之间取得平衡。虽然更高的分辨率可能编码更精细的细节，但也会增加捕捉无关因素（包括人口统计属性）的风险。研究表明，所选维度足以在自监督和对比音频表征学习中提供与说话人相关的信息，同时避免不必要的冗余 [31, 49]。我们对数梅尔频谱图按话语进行均值—方差归一化，以减少通道和响度差异。这样可以避免模型依赖绝对能量统计量，并促使其关注相对频谱模式；这在对比学习场景中尤为重要，因为录音条件否则可能会主导表征空间 [29, 30]。训练期间，我们将话语随机裁剪或用零填充至固定的 6 秒时长。这一时长既能提供足够的语音多样性以捕捉说话人特征，又能保持较高的计算效率。随机裁剪还可作为数据增强，增强模型对时间上下文的鲁棒性，并避免对特定话语片段过拟合 [30, 47]。

3.3 训练流程

3.3.1 SimCLR 基线。 我们以自监督对比学习方法 SimCLR 作为训练流程的起点。SimCLR 为研究人口统计信息泄漏提供了一个自然的基线：其训练目标明确鼓励模型对信号层面的干扰因素保持不变性，但并未纳入任何人口统计信息或公平性约束。因此，学习得到的嵌入中仍然存在的、可通过线性或非线性方式提取的任何人口统计属性，都是数据相关性隐式作用的结果，而非学习目标本身所致。理解这些嵌入的结构至关重要，因为这有助于解释：即使训练目标没有显式编码人口统计信息，这些信息为何仍可能持续存在；以及为何贸然移除这类信息可能会损害说话人判别特征。

说话人嵌入是高维向量，每个维度都编码了声学、语音和说话者特有属性的混合信息。实际上，这些嵌入彼此纠缠：某些方向主要捕捉说话人身份，另一些方向则可能无意中编码性别、年龄或口音等人口统计因素。重要的是，这些信息来源并非完全正交，也就是说，能够表征说话人身份的方向往往与承载人口统计线索的方向有所重叠。

遵循标准的 SimCLR 范式 [35]. , 通过随机波形级增强为每段输入音频生成两个相关视图。每个视图均通过独立的随机时间裁剪或零填充生成，并结合加性高斯噪声和随机振幅缩放。这些增强方式旨在扰乱与说话者身份无关的信号特征（例如录音增益和背景噪声），同时不改变底层说话者特征，从而鼓励编码器关注在此类扰动下保持稳定的特征。此处未采用时间扭曲或音高移位等其他增强方式，因为它们可能引入与人口统计特征相关性更强的变化，干扰对信息泄漏的研究。

为优化对比目标，投影头（一个包含两个隐藏层和 ReLU 激活函数的小型多层感知机）将 z 映射到一个低维对比空间，并在其中应用 NT-Xent（归一化温度缩放交叉熵）损失。投影头与编码器联合训练，但在下游评估时会被舍弃，因为先前研究表明，将编码器嵌入 z 与投影参数分离后，其在下游任务上的泛化能力更强。将投影后的表征归一化为单位长度能够稳定训练，并与评估时使用的余弦相似度保持一致。

使用 Adam 优化器以 10^{-4} 的学习率进行端到端训练。选择该学习率是因为它能兼顾稳定收敛与对梯度信号的响应能力。批量大小和温度参数 τ 依据对比学习的最佳实践设定，以确保每个批次包含足够数量的负样本对；根据经验，这有助于提升表征质量 [35]。训练期间，编码器和投影头的所有参数均可训练；没有任何组件被冻结。

训练完成后，移除投影头。使用学习到的编码器嵌入评估说话人验证和人口统计信息泄漏任务。说话人验证采用嵌入对之间的余弦相似度，因为这是一种广受认可的评分指标，能够直接反映说话人表征之间的相对接近程度，且无需复杂的后端。我们报告 ROC-AUC 和等错误率（EER）：ROC-AUC 衡量所有阈值下的可分性，而 EER 则用单个数值概括错误接受率与错误拒绝率之间的平衡情况。对于人口统计信息泄漏任务，我们通过随机梯度下降训练线性探针，以预测性别、年龄组和口音；准确率越高，表示线性可分性越强。为捕捉非线性信息，我们还评估了小型多层感知机探针，它们能够提取线性不可分的人口统计线索。

3.3.2 对抗去偏。为减轻 SimCLR 嵌入中的人口统计信息泄漏，我们引入了一种对抗去偏框架，明确抑制编码器对敏感属性进行编码。对于每种属性（性别、年龄和口音），我们都使用一个具有独立参数的对抗分类器，分别训练其从共享的编码器嵌入中预测相应属性。每个对抗器都是一个小型分类器，以编码器的嵌入为输入，并预测相应属性（例如性别），在训练过程中充当探针，用于检测表征中是否存在人口统计信息。

在编码器与每个对抗器之间应用一个梯度反转层（*GRL*）。在前向传播过程中，GRL 执行恒等映射，使对抗器能够正常接收嵌入。然而，在反向传播过程中，它会将来自对抗器的梯度乘以一个负标量，从而主动抑制编码器利用任何有助于对抗器预测人口统计属性的特征。该机制使编码器和对抗器能够通过单次端到端优化联合训练：对抗器学习更好地检测人口统计线索，同时编码器学习将其移除，而无需交替训练或多阶段训练。

每个对抗损失都由一个标量超参数 λ加权，该参数控制去偏的强度。具体而言，λ 乘以对抗损失项，用于缩放其梯度贡献，使其相对于主对比损失（NT-Xent）进行调整。较大的 λ 会施加更强的压力以去除人口统计线索，而较小的 λ 则允许编码器优先保留说话人判别结构。取值比 1.0 大几个数量级可能导致对抗梯度主导优化过程，进而造成不稳定，并使说话人特有信息丢失；这种现象在对抗式和域不变表示学习中广为人知 [36, 37]。遵循已有实践，我们因此在 1.0 附近的有限范围内对 λ进行搜索：0.2, 0.5,1.0, 2.0, 5.0。

总训练损失由 NT-Xent 对比损失和对抗损失组成，后者通过对性别、年龄和口音预测的交叉熵损失求和得到。所有可训练组件均使用 Adam 优化器进行优化，该优化器根据梯度大小及其变化性的运行估计值，自动调整每个参数的学习率。直观而言，对于梯度噪声较大或数值较大的参数，Adam 会采取更小的步长

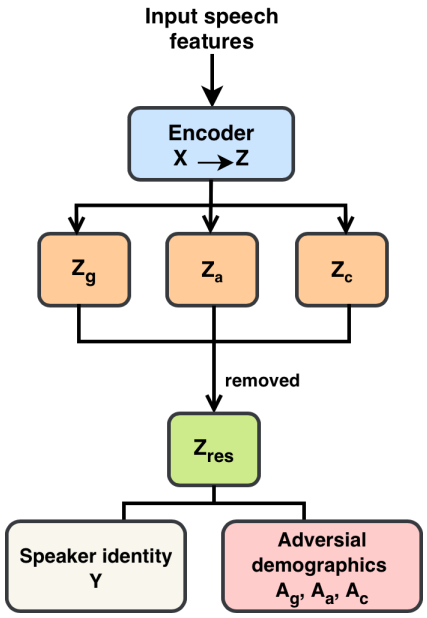


图1：因果瓶颈层架构的有向无环图（DAG）。该图展示了模型如何通过在嵌入空间中强制施加因果结构，明确地将说话人判别信息与人口统计因素（性别、年龄、口音）分离。

而对于更加稳定的参数，则采取更大的步长。这种自适应性在我们的设置中尤为有益，因为编码器同时接收来自对比目标和多个对抗目标的梯度，而这些梯度可能指向相互竞争的方向；Adam 有助于平衡这些信号，并在无需大量手动调参的情况下稳定训练。

总体而言，该设置使编码器能够保留通过 SimCLR 学习到的说话者特有特征，同时主动最小化可通过线性方式提取的人口统计信息，并且还便于后续比较添加因果瓶颈层后的效果。为评估去偏的有效性，我们在事后应用线性与非线性探针。

3.3.3 因果瓶颈层。 为了明确地将表示划分为人口统计特征子空间和说话者相关子空间，我们引入了一个因果瓶颈，将编码器输出划分为两个分支：人口统计信息分支旨在捕获显式的人口统计信息，残差分支则旨在承载独立于人口统计特征的说话者判别特征。两个分支均通过从编码器嵌入 z进行线性投影来实现。人口统计信息分支的输出 z_{demo} 通过监督学习以包含人口统计线索，而残差分支的输出 z_{res} 则被鼓励对这些线索保持不变。

对抗式梯度反转层设置应用于残差分支，并使用独立的性别、年龄和口音对抗器来抑制人口统计信息。分支之间的独立性通过协方差惩罚得到进一步促进，该惩罚最小化批次中 z_{demo} 与 z_{res} 之间的协方差。从概念上讲，这种显式解耦有助于分离因果因素：人口统计

属性在人口统计信息分支中得到更清晰的表示，而残差分支则聚焦于说话者特有的变化。

人口统计信息分支的维度 k 在 $\{32, 64, 76, 88, 100\}$之间进行扫描。较小的 k 值限制瓶颈表征人口统计因素的能力，迫使模型对这类信息进行高度压缩，这可能会增加 z_{res}中的信息泄漏。较大的 k 值允许更丰富的人口统计信息表征，但可能会减少可用于说话人判别特征的残差容量。鉴于说话人验证文献中常见的嵌入维度范围，我们选择这些取值范围来系统地探索这一权衡。瓶颈残差分支的对抗权重超参数按以下网格取值：$\{(0.01, 0.01, 0.01), (0.1, 0.01, 0.01), (0.5, 0.05, 0.05)\}$，分别对应 $\{\lambda_{\text{gender}}, \lambda_{\text{age}}, \lambda_{\text{accent}}\}$。这些取值代表轻度、中度和平衡的去偏置机制，既反映了说话人嵌入中人口统计信号的实证强度（例如，性别通常比年龄或口音更容易通过线性方式区分），也体现了在对抗压力与任务性能之间取得平衡的实际需求。远大于 1 的对抗权重可能会使优化过程不稳定，或压制说话人可判别性；在对抗式学习和域不变表示学习中，这一现象已得到广泛观察 [36, 37]。通过围绕较小到中等的λ 取值跨数量级进行扫描，我们系统地探索了从弱到强的去偏机制，同时保持联合优化的稳定性。

训练采用 Adam 进行端到端优化，学习率为10^{-3}，考虑到额外的分支参数和协方差惩罚，选择该学习率以支持更激进的更新。所有组件——编码器、人口统计信息分支、残差分支和对抗器——均参与训练。评估时会丢弃梯度反转层和对抗器，仅使用 z_{res} 计算说话人验证指标。训练完成后提取两个分支的嵌入，并通过线性与非线性探针进行泄漏评估。调整 k 和 λ，可以观察瓶颈大小和对抗强度如何共同影响人口统计信息泄漏与说话人性能，从而在查看结果前形成清晰的预测直觉。

3.4 外部数据集验证

为了进行验证，我们使用一个外部数据集，并在该数据集上测试代码。我们使用 Sonos语音控制偏差评估数据集，该数据集收录了旨在体现人口统计多样性的语音助手请求。它包含来自八个方言群体的说话者的音频样本：亚洲口音、美国西部口音、美国南部口音、美国中部口音、美国内陆北部口音、美国中大西洋地区口音、美国新英格兰口音和拉丁裔口音。该数据集还涵盖多个年龄组，其中说话者主要分布在 29–41 岁和 17–28 岁年龄段，其次是 9–16 岁、42–54 岁和 55–100 岁。每个样本都标注了说话者元数据，包括说话人标识、性别、年龄组和方言地区，因此适合用于研究人口统计偏差和训练说话者感知模型。为了使该数据集能够与 Common Voice 数据集进行比较，我们也将年龄组划分为青年（9 至 28 岁）、成年人（29 至 54 岁）和老年人（55 至 100 岁）。独立说话者共 1024 名，少于 Common Voice 数据集中的数量。我们对该数据集进行了与 Common Voice 数据集相同的分析，并比较了结果。

4 结果

本节介绍本研究的实证结果。我们首先评估 SimCLR 基线，以确定说话人验证性能，并使用线性与非线性探针量化人口统计信息泄漏。随后，我们分析对抗式去偏和因果瓶颈架构对人口统计信息泄漏及说话人验证准确率的影响，并突出各缓解策略带来的权衡。最后，我们在外部数据集上评估这些发现的鲁棒性。

4.1 SimCLR 基线

人口统计探测结果汇总于表2。仅凭嵌入即可高度准确地预测性别：测试集上的线性探针准确率超过 90%，表明与性别相关的信息具有很强的线性可分性。年龄预测则难得多：线性探针的准确率接近随机水平，而多层感知机探针的表现显著更好，这表明与年龄相关的线索是以非线性方式编码的。口音预测的整体表现最低，线性和非线性探针的准确率都相对较低，尤其是在测试集上，表明口音信息泄漏有限且不稳定。表3根据对抗去偏强度 λ报告了 Common Voice 数据集上的说话人验证性能。随着 λ 增加，性能持续下降，表明人口统计属性抑制压力与验证效用之间存在明显权衡。

总体而言，这些结果表明，尽管嵌入表示对说话人验证很有效，但也以不同程度编码了人口统计属性。性别信息被强烈且线性地嵌入，而年龄和口音的表征较弱，主要通过非线性结构体现。由于未进行重采样或平衡处理，因此这些结果应被视为人口统计信息泄漏程度的上界，而非人口统计分类的最佳表现。

表 2：Common Voice 数据集验证集和测试集上的人口统计探测结果。

属性	划分	线性探针	多层感知机探针
性别	Val	0.9120 (0.9097–0.9142)	0.9352 ± 0.0074
	Test	0.8965（0.8939–0.8991）	0.9102 ± 0.0095
Age	Val	0.4338（0.4301–0.4377）	0.4013 ± 0.0066
	Test	0.3790（0.3751–0.3829）	0.4475 ± 0.0044
口音	Val	0.2185（0.2148–0.2216）	0.2089 ± 0.0305
	Test	0.2628（0.2589–0.2664）	0.3028 ± 0.0177

4.2 对抗去偏的效果

通过在 SimCLR 对比学习基线上加入针对性别、年龄和口音的对抗分类器，并通过梯度反转进行训练，对抗去偏得到了评估。对抗信号的强度由权重参数 $\lambda_{\text{adv}} \in \{0.2, 0.5, 1.0, 2.0, 5.0\}$ 控制，从而能够系统分析说话人验证性能与人口统计信息不变性之间的权衡（见表3和表4）。对抗

去偏对人口统计信息泄漏的影响在表4中进行了量化，该表报告了不同对抗强度下线性探针和非线性探针的准确率 λ。

表 3：Common Voice 数据集上说话人验证性能随对抗强度变化的情况 λ。

λ	ROC-AUC	EER
0.2	0.8295	0.2382
0.5	0.7753	0.2922
1.0	0.7430	0.3108
2.0	0.7344	0.3212
5.0	0.6206	0.4006

随着对抗去偏强度（λ_{adv}）增加，说话人验证性能与人口统计信息泄漏之间呈现出明显的权衡。在较低的 λ （0.2–0.5）下，说话人验证性能仍然较强（ROC-AUC 为 0.868 → 0.775，等错误率为 0.202 → 0.292），而线性探针显示性别可预测性有所下降。随着λ 增加，线性探针对性别和口音的预测准确率显著下降，这表明通过因果瓶颈和对抗训练能够有效去除人口统计信息。然而，这也会牺牲说话人验证性能：在较高的 λ 下，性能会下降（ λ=5.0 时 ROC-AUC 为 0.621，等错误率为 0.401）。在所有λ下，年龄和口音预测值仍然相对较低。非线性（多层感知机）探针表明，即使在强去偏条件下，仍有部分人口统计信息残留，这凸显了对抗过滤主要针对可由线性方式获取的信息泄漏。总体而言，这些结果表明，因果瓶颈与对抗去偏相结合已经能够减少人口统计信息泄漏，但仍需仔细调整 λ，以平衡公平性和说话人验证准确率。

4.3 因果瓶颈层的影响

在所有配置下，说话人验证性能都接近随机水平，ROC-AUC 值约为 0.47–0.61（见表 5），相应的等错误率也较高。这是意料之中的，因为编码器被冻结，而残差表示经过专门训练，会丢弃与说话者无关但可用于人口统计预测的线索；与基线说话人嵌入相比，这显著削弱了身份可分性。我们观察到，随着 k 值和对抗强度变化（见表 6 和表 7 ），结果出现中等程度的波动；这可能是因为在线性瓶颈必须使用固定的编码器特征时，需要在人口统计集中与残差抑制之间进行权衡。人口统计信息分支中的性别准确率始终低于基线（0.90–0.94），这反映出性别信息不再自由地编码于整个表征中，而是被限制在一个低维线性瓶颈中；该瓶颈还受到正交性和对抗目标的制约。即使在受保护分支中，这种有意的限制也降低了绝对可预测性。相比之下，人口统计信息分支中的年龄和口音准确率均高于各自的基线，这是因为显式监督会将这些属性推入专用子空间，从而有效集中那些此前在基线嵌入中编码得较弱且分散的信息。至关重要的是，对于所有

表 4：在 Common Voice 数据集上，人口统计属性的线性探针和多层感知机探针准确率随对抗强度变化的情况 λ。线性探针报告准确率及其 95% 置信区间；多层感知机探针报告不同随机种子下的均值 ± 标准差。

属性	λ	线性准确率（置信区间）	多层感知机准确率（± 标准差）
性别	0.2	0.9120 [0.9097–0.9142]	0.9352 ± 0.0074
	0.5	0.8965 [0.8939–0.8991]	0.9102 ± 0.0095
	1.0	0.8542 [0.8510–0.8575]	0.7766 ± 0.0080
	2.0	0.8086 [0.8052–0.8119]	0.7645 ± 0.0072
	5.0	0.6749 [0.6700–0.6798]	0.6999 ± 0.0065
Age	0.2	0.3685 [0.3650–0.3720]	0.4699 ± 0.0050
	0.5	0.4108 [0.4070–0.4146]	0.4753 ± 0.0062
	1.0	0.3818 [0.3780–0.3856]	0.4662 ± 0.0058
	2.0	0.3797 [0.3760–0.3835]	0.4750 ± 0.0055
	5.0	0.3743 [0.3705–0.3781]	0.4725 ± 0.0061
口音	0.2	0.2320 [0.2280–0.2360]	0.4087 ± 0.0071
	0.5	0.2195 [0.2155–0.2235]	0.4117 ± 0.0068
	1.0	0.2247 [0.2207–0.2287]	0.4122 ± 0.0070
	2.0	0.2355 [0.2315–0.2395]	0.4119 ± 0.0065
	5.0	0.1585 [0.1545–0.1625]	0.4112 ± 0.0062

表 5：各个k 值对应的 ROC-AUC 和等错误率（EER）得分，展示模型的整体验证性能。每行列出训练期间使用的性别、年龄和口音对应的 lambda 权重（λ_{gender}， λ_{age}， λ_{accent}）。此表展示了不同的属性正则化组合如何影响模型在不同 k设置下的判别能力，评估指标为 ROC-AUC 和 EER。

k	$\lambda_{gender}, \lambda_{age}, \lambda_{accent}$	ROC AUC	EER
32	0.01, 0.01, 0.01	0.3940	0.6300
32	0.1, 0.01, 0.01	0.5300	0.4603
32	0.5, 0.05, 0.05	0.4530	0.6895
64	0.01, 0.01, 0.01	0.3888	0.5962
64	0.1, 0.01, 0.01	0.5967	0.4097
64	0.5, 0.05, 0.05	0.4501	0.4881
76	0.01, 0.01, 0.01	0.5340	0.6329
76	0.1, 0.01, 0.01	0.5215	0.5764
76	0.5, 0.05, 0.05	0.6062	0.5347
88	0.01, 0.01, 0.01	0.6254	0.4891
88	0.1, 0.01, 0.01	0.4822	0.6300

在三个属性上，残差分支的准确率均持续低于人口统计信息分支，且在若干情况下低于基线准确率，尤其是在年龄和口音方面。这表明，对抗目标成功抑制了残差表示中的人口统计信息泄漏，这也是所提方法的主要目标。表7报告了不同瓶颈尺寸下残差分支的人口属性预测准确率

表6：人口统计信息分支中性别、年龄和口音的人口统计属性线性分类器测试准确率。准确率按 k报告。

k	性别	Age	口音
32	0.704	0.555	0.381
64	0.695	0.556	0.381
76	0.695	0.556	0.382
88	0.696	0.555	0.381
100	0.676	0.386	0.247

表7：不同 k值下，残差分支对性别、年龄和口音预测的准确率得分。每行显示仅残差分支的测试准确率，以及用于性别、年龄和口音的相应 λ 值（λ_{gender}、 λ_{age}、 λ_{accent}）。

k	λ_{gender}、 λ_{age}、 λ_{accent}	残差分支性别	残差年龄	残差口音
32	0.01, 0.01, 0.01	0.6787	0.4385	0.3809
32	0.1、0.01、0.01	0.6992	0.3867	0.1406
32	0.5, 0.05, 0.05	0.6895	0.3867	0.1406
64	0.01, 0.01, 0.01	0.5849	0.4141	0.1563
64	0.1, 0.01, 0.01	0.6807	0.3389	0.2025
64	0.5, 0.05, 0.05	0.6660	0.3320	0.1932
76	0.01, 0.01, 0.01	0.5820	0.3955	0.3359
76	0.1, 0.01, 0.01	0.6699	0.3955	0.2766
76	0.5, 0.05, 0.05	0.6172	0.3613	0.1932
88	0.01, 0.01, 0.01	0.6504	0.3125	0.3809
88	0.1, 0.01, 0.01	0.6904	0.3125	0.3809

以及不同对抗权重设置下的人口属性预测准确率，展示了因果分离后人口统计信息被抑制的程度。

4.4 外部验证

4.4.1 SimCLR 对比。为进行验证，我们使用 Sonos 数据集作为外部验证数据集。尽管与 Mozilla Common Voice 相比，绝对说话者验证性能有所下降（这可能是由于独立说话者数量较少），但总体趋势保持一致（见表 8 和表 9）。在两个数据集中，性别信息都具有很高的可预测性，而年龄预测仍然较弱且不稳定。口音预测表现出明显的数据集依赖性，在 Sonos 上的性能骤降至接近随机水平。重要的是，人口统计特征可预测性的相对排序（性别 > 年龄 > 口音）在不同数据集中保持不变，表明这些模式是表征学习固有的。

4.4.2 对抗去偏比较。 表10报告了 Sonos 数据集上说话人验证性能随对抗正则化强度 λ变化的情况，从而可以直接比较去偏压力不断增大时验证性能的下降。随着对抗正则化强度 λ_{adv} 增加，两个数据集上的说话人验证性能均有所下降，同时性别可预测性降低，表明去偏效果显著。年龄预测仍然较弱，且基本不受 λ_{adv}影响，这表明学到的表示中年龄相关信息有限

表8：外部验证 Sonos 数据集上的总体及各子组说话人验证指标。指标包括 ROC-AUC 和等错误率。

组别	ROC-AUC	EER
总体		
	0.7371	0.3301
性别		
女性	0.6500	0.3930
Male	0.7410	0.3270
年龄组		
青年	0.7170	0.3490
成年人	0.7570	0.3120
老年人	0.7060	0.3570
口音组		
亚洲口音	0.7360	0.3310
拉丁裔口音	0.7520	0.3100
美国内陆北部口音	0.7950	0.2810
美国中大西洋地区口音	0.7390	0.3270
美国中部口音	0.6800	0.3730
美国新英格兰口音	0.8630	0.2280
美国南部口音	0.7440	0.3290
美国西部口音	0.7370	0.3350

表9：Sonos 数据集验证集和测试集划分上的线性探针与多层感知机探针准确率（%）。线性探针的置信区间（CI）列于方括号中。多层感知机探针报告了不同随机种子下的均值± 标准差。

属性	划分	线性准确率（置信区间）	多层感知机准确率（均值 ± 标准差）
性别	Val	0.7784 [0.7648, 0.7914]	0.7899 ± 0.0021
	Test	0.8174 [0.8074, 0.8267]	0.8267 ± 0.0029
Age	Val	0.4819 [0.4659, 0.4967]	0.5048 ± 0.0540
	Test	0.3745 [0.3619, 0.3881]	0.4023 ± 0.0443
方言	Val	0.1466 [0.1363, 0.1576]	0.1218 ± 0.0075
	Test	0.1157 [0.1075, 0.1241]	0.1152 ± 0.0061

表10：偏差正则化强度 λ 对 Sonos 数据集上说话人验证性能的影响。

λ	ROC-AUC	EER
0.2	0.6389	0.4170
0.5	0.6045	0.4302
1.0	0.5951	0.4275
2.0	0.5797	0.4357
5.0	0.5630	0.4635

表征。口音预测对数据集的依赖性更强：虽然对抗去偏降低了主数据集上的口音泄漏，但 Sonos 数据集上的口音准确率本就已接近随机水平，对 λ_{adv} 的敏感度也有限。总体而言，这些权衡在不同数据集上均得以保持，支持了所提出去偏方法的鲁棒性。表11总结了 Sonos 数据集上不同 λ_{adv} 取值下的线性和非线性人口统计探测结果，

表11：外部 Sonos 数据集上，人口统计属性的线性探针和多层感知机探针准确率随偏差正则化强度 λ 变化的结果。线性探针报告带有 95% 置信区间的准确率；多层感知机探针报告跨随机种子的均值 ±标准差。

属性	λ	线性准确率（置信区间）	多层感知机准确率（±标准差）
性别	0.2	0.6183 [0.6044–0.6305]	0.5343 ± 0.0000
	0.5	0.5995 [0.5856–0.6119]	0.5343 ± 0.0000
	1.0	0.5625 [0.5488–0.5747]	0.5343 ± 0.0000
	2.0	0.5415 [0.5288–0.5554]	0.5343 ± 0.0000
	5.0	0.5321 [0.5188–0.5456]	0.5343 ± 0.0000
Age	0.2	0.3770 [0.3635–0.3899]	0.3579 ± 0.0000
	0.5	0.4140 [0.4009–0.4280]	0.3579 ± 0.0000
	1.0	0.3772 [0.3646–0.3898]	0.3579 ± 0.0000
	2.0	0.3566 [0.3438–0.3697]	0.3579 ± 0.0000
	5.0	0.3380 [0.3263–0.3515]	0.3593 ± 0.0011
口音	0.2	0.1305 [0.1212–0.1396]	0.1662 ± 0.0184
	0.5	0.1691 [0.1594–0.1788]	0.1693 ± 0.0249
	1.0	0.1573 [0.1474–0.1671]	0.1680 ± 0.0297
	2.0	0.1203 [0.1108–0.1290]	0.1689 ± 0.0281
	5.0	0.1292 [0.1201–0.1381]	0.1545 ± 0.0230

展示了对抗去偏在域偏移情况下对人口统计信息泄漏的影响。

5 讨论

在本研究中，我们系统地分析了自监督说话人嵌入中的人口统计信息泄漏，并评估了对抗去偏和因果瓶颈层作为缓解策略的有效性。在 Common Voice 和外部验证 Sonos 数据集上，都呈现出一致的人口统计信息可预测性层级：性别的编码最为显著，其次是年龄，而口音最难恢复。外部数据集也呈现出这一规律，表明这些模式在不同数据集之间具有一致性，反映了说话人嵌入自然编码信息的方式。值得注意的是，Sonos 数据集在口音/方言预测上的表现差得多。一种可能的解释是，即便对人类听者而言，区分不同方言也比区分不同口音更难。

性别对抗去偏的响应最强，而年龄和口音则没有明显响应，这表明对抗训练主要抑制了以线性方式编码的属性。众所周知，性别信息与主要声学维度高度相关，因此会产生稳定的对抗梯度 [44]。相比之下，年龄和口音似乎以更非线性、分布更分散的方式编码，因此更难通过标准的梯度反转将其消除。值得注意的是，尽管对抗去偏对年龄和口音的影响不大，但其影响仍足以损害说话人验证性能，这表明说话人验证性能在很大程度上依赖于性别。

引入因果瓶颈层后，在线性与非线性探针上的性别信息泄漏进一步减少，但代价是说话人验证性能大幅下降，且通常比单独使用对抗式去偏时下降得更严重。这一观察进一步支持了这样的解释：瓶颈在移除人口统计信息的同时，也移除了说话人验证所需的信息。此外，在某些配置下，人口统计信息分支在年龄和口音准确率方面优于基线嵌入。这一现象并不违背去偏目标，而是可以用信息集中来解释：显式监督会迫使编码较弱的属性集中到一个低维子空间中，即使整体表征能力有所下降，探针性能仍会提升。相比之下，人口统计信息分支的性别准确率始终低于基线。这是因为性别信息在原始嵌入中已经得到了强烈且线性的编码；引入瓶颈会限制这一主导信号，而非将其集中起来，从而削弱主要的性别方向，并使探针准确率低于不受约束的基线。同时，残差分支的人口统计准确率始终低于人口统计信息分支，表明因果分离目标取得了部分成效。

减少人口统计信息泄漏最成功的策略是因果瓶颈层，其效果显著优于对抗去偏。残差分支始终能将人口统计信息泄漏降低至基线以下。在线性探针中，性别信息从 90% 降至 57%，降幅显著。年龄泄漏从 38% 降至 27%，口音泄漏从 26% 降至 18%。然而，我们发现这种降低的代价非常高，因为 ROC-AUC 从 82% 降至约 50%。因此，这一策略清晰地体现了效用—公平性权衡。

5.0.1 与既有工作的比较。我们关于性别可以从说话人嵌入中被高度预测的发现，与既有关于预训练语音模型中人口统计信息泄漏的研究高度一致。例如，[45] 报告称，大规模预训练编码器的性别分类准确率超过 99%，这表明性别信息在现代语音表征中既十分显著，又可以通过线性方式轻易获取。尽管我们的基线 SimCLR 嵌入表现出较低的绝对性别准确率（约为 90%—94%），但鉴于模型容量较小且缺乏显式的说话人监督，这一结果是预期之中的。尽管如此，定性结论是一致的：性别与音高范围和共振峰结构等说话人判别性声学特征深度交织。

我们的实验中，年龄预测表现中等，优于以往研究所预期的水平；以往研究将年龄估计视为一项困难的回归任务，其平均绝对误差可达数年 [45]。这种表面上的差异可以用我们采用粗略的年龄组标签，而非细粒度的连续年龄来解释。基于年龄组的预测大幅降低了任务复杂度，也可能使年龄相关线索（如声音成熟度或衰老）更加明显。所有配置下的口音预测表现仍然较弱，这与先前的观察一致：口音信息的编码分布更分散，且高度依赖数据集构成、说话人重叠和标签噪声。Sonos 数据集上的口音可预测性崩溃

进一步支持了这样的解释：口音线索并非说话人身份的内在特征，而是更加依赖特定数据集分布。

通过梯度反转进行对抗去偏成功减少了性别信息泄漏，尤其是对线性探针而言，但对年龄和口音可预测性的影响有限。这一表现与说话人识别公平性领域先前研究的发现一致：对抗目标往往能改善群体公平性指标，但会显著损害任务效用 [42]。在我们的实验中，增大对抗强度 λ_{adv} 始终会降低说话人验证性能。

6 结论

本文考察了自监督说话人嵌入中的人口统计信息泄漏，并评估了对抗去偏和因果瓶颈架构能否在不显著降低说话人验证性能的情况下缓解此类泄漏。我们使用 SimCLR 训练的嵌入进行研究，结果表明，不同属性的编码程度各不相同：性别可以通过线性方式强有力地恢复，年龄的编码较弱且通常为非线性，口音则最难预测。非线性探针持续揭示出线性探针未能发现的额外信息泄漏，尤其是年龄方面，这证实仅靠线性评估会低估人口统计信息含量。

对抗去偏减少了性别信息泄漏，但对年龄和口音的效果较差，尤其是在非线性探测下。进一步提高对抗强度能够抑制人口统计属性的可预测性，但也导致说话人验证性能显著下降，凸显出公平性与效用之间明显的权衡。引入因果瓶颈后，可以更明确地区分人口统计信息与残差信息，并且在若干情况下，使残差表示中的人口统计属性可预测性低于基线水平。然而，这也付出了说话人验证性能大幅下降的代价，且未能始终消除非线性人口统计信息泄漏，表明人口统计属性仍与说话人判别特征纠缠在一起。

总体而言，结果表明，自监督 d
说话人嵌入中的人口统计信息泄漏可以减少，但若不牺牲效用则无法彻底消除。相较于以往工作，本文系统比较了对抗去偏和因果去偏下的线性与非线性探针，并证明尽管因果瓶颈能够更有力地控制信息泄漏，却会加剧公平性与性能之间的权衡。这些结论受到若干因素的限制，包括将探测分类器用作代理测量指标、人口统计标签较为粗略，以及使用冻结编码器。未来工作应探索能够显式针对非线性结构的去偏目标，并评估能否在对说话人验证性能影响更小的情况下实现更强的不变性。

参考文献

[1] B人工智能生成的深度伪造与错误信息激增：21世纪全球关注的威胁。*Qeios*，2023年12月。doi：10.32388/DPLE2L。[2] Facebook。面向政府和非营利组织的错误信息资源，2025年。https://www.facebook.com/government-nonprofits/blog/misinformation-resources。 [访问日期：12-Feb-2025]。[3] X（原 Twitter）。更新我们应对误导性信息的方法，2020年。https://blog.x.com/en_us/topics/product/2020/updating-our-approach- to-misleading-information。 [访问日期：12-Feb-2025]。

[4] J. S. Brennen、F. M. Simon 和 R. K. Nielsen。《超越（错误）表征：公共危机中的错误信息》。Journalism Studies，23(5–6):677–694，2022。doi: 10.1080/1461670X.2022.2031259. https://www.tandfonline.com/doi/full/10.1080/1461670X.2022.2031259. [访问日期：12-Feb-2025]。[5] 马克·扎克伯格。是时候回归我们言论自由的初心了：我们将以其他方式取代事实核查，2025。https://www.facebook.com/zuck/videos/its-time-to-get-back-to-our-roots-around-free-expression-were-replacing-fact-che/1525382954801931/. [访问日期：12-Feb-2025]。[6] X。社区笔记，2025。https://help.x.com/en/using-x/community-notes. [访问日期：12-Feb-2025]。[7] Y. Chuai、H. Tian、N. Pröllochs 和 G. Lenzini。社区笔记上线后，X/Twitter 上错误信息的互动量是否有所下降？arXiv，2023。doi: 10.48550/arXiv.2307.07960. https://arxiv.org/abs/2307.07960. [访问日期：12-Feb-2025]。[8] Y. Chuai、M. Pilarski、G. Lenzini 和 N. Pröllochs。社区笔记减少了 X 上误导性帖文的传播。arXiv，2024。https://arxiv.org/abs/2409.08781. [9] C. P. Drolsbach、K. Solovev 和 N. Pröllochs。社区笔记减少了 X 上误导性帖文的传播。PNAS Nexus，2024。https://doi.org/10.1093/ pnasnexus/pgae217. [访问日期：12-Feb-2025]。[10] H. Habib、S. Elsharawy 和 R. Rahman。“我不信任他们”：探索人们对标记网络错误信息的事实核查机构的看法。arXiv，2024。arXiv:2410.00866. https://arxiv.org/abs/2410.00866. [11] S. Park、J. Y. Park、H. Chin、J.-h. Kang 和 M. Cha。理解事实核查消息引发的用户体验与感知偏差：一项实验研究。收录于《2021年万维网大会论文集》2780，2021 21 2769 10 1145 3442381 3450121 10 1145 3442381 3450121 (WWW ')，页码 – 。doi: . / . . https://doi.org/ . / . . [12] X. Liu、L. Qi、L. Wang 和 M. J. Metzger。核查事实核查者：信源类型、感知可信度及个体差异在事实核查有效性中的作用。Communication Research，2023。doi: 10.1177/00936502231206419. https://doi.org/10.1177/00936502231206419. [13] J. Li 和 X. Chang。通过分享真相打击错误信息：社交媒体上事实核查内容传播研究。Information Systems Frontiers，25:1479–1493，2023。doi: 10.1007/s10796-022-10296-z. https://doi.org/10.1007/s10796-022- 10296-z. [访问日期：12-Feb-2025]。[14] U. Kangur、R. Chakraborty 和 R. Sharma。谁来核查核查者？探索 Twitter 社区笔记的信源可信度。arXiv，2024。arXiv:2406.12444. https://arxiv.org/abs/2406.12444. [15] N. Pröllochs。Twitter 的 Birdwatch 平台上的社区事实核查。收录于《国际 AAAI 网络与社交媒体会议论文集》，16(1):794–805，2022。doi: 10.1609/icwsm.v16i1.19335. https://doi.org/10.1609/icwsm.v16i1.19335.[16] F. Panizza、P. Ronzani、T. Morisseau 等。网络用户如何回应众包事实核查？Humanities and Social Sciences Communications，10:867，2023。doi: 10.1057/s41599-023-02329-y. https://doi.org/10.1057/s41599-023-02329-y. [17] T. Bucher。危机中的审核？收购之后社交媒体内容审核实践的变化。Social Media + Society，9(2)，2023。doi: 10.1177/20563051231121720。[18] S. Frenkel。Twitter 内容审核的衰退：马斯克收购后的变化如何影响内容管控。《纽约时报》，2023。[19] R. Gorwa。平台治理三角：阐释国家、企业与公民社会在数字媒体平台监管中的作用。Internet Policy Review，8(1)，2019。[20] T. Gillespie。《互联网的守护者：平台、内容审核以及塑造社交媒体的隐秘决策》。耶鲁大学出版社，2018。[21] N. Suzor。《无法无天：支配我们数字生活的隐秘规则》。剑桥大学出版社，2019。[22] C. Newton。《创伤楼层：美国 Facebook 内容审核员的隐秘生活》。The Verge，2020。https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona. [23] S. Bradshaw 和 P. N. Howard。《全球虚假信息秩序：2019年有组织社交媒体操纵全球清单》。牛津互联网研究所，2019。doi: 10.14763/2019.1.1413。[24] K. Klonick。新任治理者：规范网络言论的人们、规则与流程。《哈佛法律评论》，131(6):1598–1670，2018。[25] S. Zannettou、M. Sirivianos、J. Blackburn 和 N. Kourtellis。虚假信息网络：社交媒体上的谣言与错误信息动态。Proceedings of the ACM on Human-Computer Interaction，4(CSCW2):1–26，2020。doi: 10.1145/3415178。[26] Z. Tufekci。《Twitter 与催泪瓦斯：网络化抗议的力量与脆弱性》。耶鲁大学出版社，2018。[27] E. Douek。将内容审核视为系统思维。Knight First Amendment Institute，2020。https://knightcolumbia.org/content/content-moderation- as-systems-thinking. [28] R. Caplan 和 T. Gillespie。分层治理与取消变现：平台经济中劳动与报酬条款的变化。Social Media + Society，6(2)，2020。doi: 10.1177/2056305120936636。

[29] D. Snyder、D. Garcia-Romero、G. Sell、D. Povey 和 S. Khudanpur。X向量：用于说话人识别的稳健深度神经网络嵌入。收录于 IEEE国际声学、语音与信号处理会议论文集，页码 5329–5333，2018。[30] B. Desplanques、J. Thienpondt 和 K. Demuynck。ECAPA-TDNN：在基于时延神经网络的说话人验证中强调通道注意力、传播与聚合。收录于 Interspeech 论文集，页码 3830–3834，2020。doi: 10.21437/Interspeech.2020- 2650。[31] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli。wav2vec 2.0：语音表征自监督学习框架。收录于《神经信息处理系统大会进展》（NeurIPS），2020。[32] D. Raj、D. Snyder、D. Povey 和 S. Khudanpur。探究 x 向量中编码的信息。收录于 Interspeech 论文集，页码 2943–2947，2019。doi: 10.21437/Interspeech.2019-2169。[33] L. Luu、S. Renals 和 P. Bell。利用解耦说话人表示研究说话人属性对说话人可分性的贡献。收录于 Interspeech 论文集，页码 3068–3072，2022。doi: 10.21437/Interspeech.2022-10567。[34] M. J. Kusner、J. R. Loftus、C. Russell 和 R. Silva。反事实公平性。arXiv 预印本 arXiv:1703.06856，2017。https://arxiv.org/abs/1703.06856. [35] T. Chen、S. Kornblith、M. Norouzi 和 G. Hinton。视觉表示对比学习的简单框架。arXiv 预印本 arXiv:2002 05709，2020 2002 05709 . . https://arxiv.org/abs/ . . [36] Y. Ganin、E. Ustinova、H. Ajakan、P. Germain、H. Larochelle、F. Laviolette、M. Marc-hand 和 V. Lempitsky。神经网络的域对抗训练。Journal of Machine Learning Research，17(59):1–35，2016。[37] Y. Elazar 和 Y. Goldberg。从文本表示中对抗性地移除人口统计属性。收录于 EMNLP 论文集，2018。[38] H. Edwards 和 A. Storkey。利用对抗者审查表征。收录于国际学习表示会议（ICLR），2016。[39] B. H. Zhang、B. Lemoine 和 M. Mitchell。利用对抗学习缓解不必要的偏差。收录于 AAAI/ACM人工智能、伦理与社会会议（AIES）论文集，页码 335–340，2018。[40] R. Peri、K. Somandepalli 和 S. Narayanan。是否进行对抗训练：说话人识别偏差缓解策略研究。Computer Speech & Language，79:101481，2022。doi: 10.1016/j.csl.2022.101481。[41] R. Peri、K. Somandepalli 和 S. Narayanan。说话人识别偏差缓解策略研究。Computer Speech & Language，79:101481，2023。doi: 10.1016/j.csl.2022.101481。[42] H. Shen、Y. Yang、G. Sun、R. Langman、E. Han、J. Droppo 和 A. Stolcke。通过群组自适应融合网络提升说话人验证的公平性。arXiv 预印本 arXiv:2202.11323，2022。https://arxiv.org/abs/2202.11323. [43] O. Baser、A. E. Tanriverdi、K. Kale、S. P. Chinchali 和 S. Vishwanath。WavShape：面向公平且注重隐私的音频处理的信息论语音表示学习。收录于 Interspeech 论文集，2025。doi: 10.21437/Interspeech.2025-2528. https://arxiv.org/abs/2506.22789. [44] X. Zhang、S. Wang、Z. Chen、Y. Qian 和 K. Yu。利用 WavLM 嵌入从语音中预测人口统计属性。arXiv 预印本 arXiv:2305 14058，2023 . .[45] Y. Yang、T. Thebaud 和 N. Dehak。利用 WavLM 嵌入从语音中预测人口统计属性。arXiv 预印本 arXiv:2502.12007，2025。https: //arxiv.org/abs/2502.12007. [46] W. T. Hutiri 和 A. Y. Ding。自动说话人识别中的偏差。收录于 ACM FAccT 论文集，页码 230–247，2022。doi: 10.1145/3531146.3533089. https://doi.org/10.1145/3531146.3533089.[47] J. S. Chung、A. Nagrani 和 A. Zisserman。VoxCeleb2：深度说话人识别。收录于 Interspeech 论文集，2018。[48] L. R. Rabiner 和 R. W. Schafer。《语音信号数字处理》。Prentice-Hall，1978。[49] D. S. Park、W. Chan、Y. Zhang、C.-C. Chiu、B. Zoph、E. D. Cubuk 和 Q. V. Le。SpecAugment：一种用于自动语音识别的简单数据增强方法。收录于 Interspeech 论文集，2019。https://arxiv.org/abs/1904.08779. [50] Mozilla基金会。Mozilla Data Collective数据集，2025。https://datacollective.mozillafoundation.org/datasets. [访问日期：2025-12-16]。[51] Z. Li、Q. Xu 和 Y. Song。句子嵌入泄露的信息超乎预期：通过生成式嵌入反演攻击还原整句。收录于 ACL 论文集精选，页码 14049–14064，2023。doi: 10.18653/v1/2023.findings- acl.881. https://aclanthology.org/2023.findings-acl.881. [52] E. Koçadağ、S. S. Mohammadi Ziabari 和 A. M. Mansoor Alsahag。评估基于结构因果模型的交叉性偏差缓解后词嵌入的质量与连贯性。arXiv 预印本 arXiv:2601.04393，2026。doi: 10.48550/arXiv.2601.04393。[53] C. Zhu、S. S. Mohammadi Ziabari 和 A. M. Mansoor Alsahag。利用结构因果模型进行情感分析的任务自适应去偏。arXiv 预印本，2025。[54] A. Garbat、S. S. Mohammadi Ziabari 和 A. M. Mansoor Alsahag。上下文嵌入的温暖感与能力感去偏。收录于第20届 IEEE 国际语义计算会议论文集，2026。[55] B. Naranbat、S. S. Mohammadi Ziabari 和 A. M. Mansoor Alsahag。使用基于Transformer的模型探索多域道德情感分类中的公平性指标设计。arXiv 预印本 arXiv:2510.11222，2025。[56] X. Li、Y.Chen 和 L. Sun。FairASR：面向自动语音识别的公平音频对比学习。发表于 *Interspeech* 论文集，2023 年。doi: 10.21437/Interspeech.2023- 1234.[57] S. Ruder、D. Jones 和 M. Chen。自动语音识别中的公平性并非一刀切。计算语言学协会会刊，10：245–259，2022 年。doi：10.1162/tacl_a_00412。[58] J. Park、H. Kim 和 S. Lee。性别不变语音表征的对比学习。发表于 *EMNLP* 论文集，2023 年。doi: 10.18653/v1/2023.findings- emnlp.87。[59] Y. Xie 等。FairIB：用于公平表征学习的信息瓶颈。发表于 国际人工智能联合会议（*IJCAI*）论文集，2024 年。