合乎伦理的语音识别数据集：相关考量

Orestis Papakyriakopoulos Sony
AI Switzerland
orestis.papakyriakopoulos@sony.com

Alice Xiang Sony
AI USA
alice.xiang
@sony.com

arXiv:2305.02081v1 [cs.CY] 3 May 2023

摘要

语音人工智能技术大多基于公开数据集或通过大规模语音网络爬取进行训练。在这两种情况下，数据采集都着重于尽量减少收集工作，却未必会考虑数据主体的保护或用户需求。因此，当这些模型用于与训练集中主导人口群体存在差异的用户时，往往不够稳健，并会歧视使用不同方言、口音、说话风格以及存在言语不流畅现象的个体。本次演讲将以自动语音识别为案例，探讨合乎伦理的语音数据集应具备哪些特性，才能推动负责任的人工智能应用。我们将展示多样性问题、包容性实践及必要的考量，这些因素既能改进训练出的模型，也有助于提升模型的可解释性，并保护用户和数据主体。我们主张依法保护数据主体的隐私，依据用户的人口统计特征和需求进行定向数据抽样，采用恰当的元数据，以确保模型失效时的可解释性与问责，并进行社会技术层面的情境化模型设计。我们希望本次演讲能启发研究人员和从业者，在语音技术及其他领域设计并使用更多以人为中心的数据集，以赋能并尊重用户，同时提升机器学习模型的稳健性和效用。

CCS 概念

• **社会与专业主题** → **用户特征**；**计算/技术政策**；• **以人为中心的计算** →**协作与社会计算设计与评估方法**。

关键词

自动语音识别，以人为本的人工智能，数据收集，算法公平性

ACM 参考格式：
Orestis Papakyriakopoulos 和 Alice Xiang。2023年。《伦理语音识别数据集的考量》。载于第十六届 *ACM* 国际网络搜索与数据挖掘会议论文集（*WSDM '23*），*2023*年*2*月*27*日至*3*月*3*日，新加坡，新加坡。ACM，美国纽约州纽约市，3页。https://doi.org/10.1145/3539597.3575793

当前版本是与 ACM WSDM'23 发表版本相对应的预印本。引用已发表版本时，请使用 ACM 参考格式。允许免费为个人或课堂使用而制作本作品的部分或全部内容的数字或纸质副本，但不得出于营利或商业利益而制作或分发副本，且副本首页必须包含此声明和完整引文。必须尊重本作品第三方组件的版权。其他用途请联系所有者/作者。*WSDM '23*，*2023*年*2*月*27*日至*3*月*3*日，新加坡，新加坡© 2023年版权归所有者/作者所有。ACM ISBN 978-1-4503-9407-9/23/02。https://doi.org/10.1145/3539597.3575793

1 数据与多样性

模型训练的数据集选择与整理是机器学习中的基础步骤。数据集不仅与模型的稳健性相关，还涉及严重的伦理问题，例如所包含个体的隐私、表征性伤害，以及下游使用中的偏见 [8]。事实上，在语音技术领域，既有研究表明，商业自动语音识别（ASR）应用会歧视使用非主流方言、口音和说话风格的个体 [5–7], , 语音识别对这些人的错误率更高。因此，我们认为，理解现有数据集的局限性十分重要，因为这些局限性会导致有问题的模型结果；同时，还应制定并采用能够实现包容性模型部署的数据收集与整理实践。

大多数自动语音识别模型要么使用开放源代码数据集[1] ，要么抓取开放网络 [9]进行训练。然而，在开发负责任的人工智能应用时，这些方法面临着严重局限。通过审视100多个开放源代码数据集的特征，我们发现，大多数数据集包含的语音都源自占主导地位的语言变体，却没有纳入现有方言或非主流社会群体所使用的口音。此外，数据集在年龄和性别多样性方面也存在严重局限：有些数据集完全没有涵盖特定性别或年龄群体，有些则极度失衡。这一特征导致基于这些数据集训练的模型在特定交叉群体 [4, 11]上的表现不佳。此外，很少有数据集包含存在言语不流畅或构音障碍的个体语音，这表明人们对能够在处于不利地位的社会群体中稳健运行的自动语音识别技术关注极少。同样重要的是，开放源代码数据集在领域特定语音方面的多样性有限，其内容往往收集自孤立且特殊的来源（例如《圣经》 [2]、议会 [10]），与特定社群的社会政治现实存在偏差。类似地，人们对录音条件的多样性（背景噪声、背景语音）以及所使用录音媒介的多样性（录音设备属性等）关注有限，而这些因素可能对自动语音识别模型的性能产生不利影响，并且与个体的社会经济地位相关。

大规模抓取在线资源被认为是解决上述许多问题的一种方法 [3, 9], , 因为这样更有可能使所用数据集涵盖更多样的口音、方言、语言和说话风格。尽管如此，我们指出，这种做法可能会造成进一步的负面影响。首先，抓取到的语音要么附带质量未知的字幕，要么需要通过自动语音识别技术转录，而这些技术会重现既有偏见；或者由不熟悉语料库中个体口音、方言和说话风格的人工标注员转录。其次，收集到的数据仍然.....

不平衡且晦涩，因为爬取语音的确切属性从未得到提供，导致模型性能不佳时出现问责和可解释性方面的问题。第三，即使在法律上允许，网络数据收集也可能侵犯个人隐私，并使其面临相关风险，而他们对此并不知情。因此，我们认为，在训练自动语音识别（ASR）技术时，有必要更加谨慎地考虑所使用的数据，以减轻对用户和数据主体造成的不必要伤害及后果。

2 最佳实践

考虑到现有数据集和数据收集实践的局限性，我们提出一系列最佳实践，旨在减少危害、包容不同的社区和用户，同时强调在开发人工智能应用时跨学科研究和以人为中心的方法的重要性。这些实践包括以下特性：

- **基于情境的数据收集**。 数据集应包含能够体现语言多样性的观察数据，涵盖性别、年龄、方言、口音、说话风格、录音环境和录音媒介，并与未来将使用所开发语音技术的人群相对应。应将这些特性纳入语料库的元数据，以便在模型失效时提高可解释性。

- **社会技术设计**。 应在明确自动内容识别应用的技术、伦理和社会政治要求后，设计数据集的特性。这意味着数据管理者需要了解用户群体的不同人口特征如何体现在数据集的各类口音、方言和说话风格中。此外，还要考虑语音语料库如何在词汇多样性和代表性方面体现并赋能用户的社会政治背景。

- **反思性转录**。 如果需要转录语音观察记录，理想情况下，标注员的人口统计特征应与说话者相符。这有助于避免错误识别词语和表达方式。如果无法做到这一点，则必须报告相关局限性以及标注员的人口统计特征。

- **知情同意、隐私与对数据主体的尊重**。数据主体应签署同意书，明确说明将使用其数据进行训练的自动语音识别应用的确切范围和背景。如果要公开数据集，相关许可协议中必须明确规定保障措施，以保护个人免遭重新识别。此外，数据管理者应实施适当机制，以便在收到请求后删除数据主体信息。数据集创建者还应根据所在国家的劳动政策，就数据主体付出的劳动给予补偿。

- **善意的模型部署**。 为负责任的人工智能应用而开发的数据集，应提供明确的许可指南，以规范模型的善意使用。数据集创建者应限制将这些数据用于生物识别或监控等恶意目的。

我们认为，以上实践能够显著提升自动语音识别技术的性能，同时推动负责任人工智能的发展。我们相信，将用户和数据主体的利益与需求置于核心位置，研究人员和从业者就能开发出更具包容性、多样性且更能赋能人们的数据集，同时提升模型的鲁棒性。我们希望，我们的报告能促使语音及更广泛的机器学习领域重新评估并反思在开发和使用模型训练数据集时所采取的实践。

3 公司简介

作为索尼集团公司的全资子公司，索尼人工智能（Sony AI）于2020年4月成立，旨在加速人工智能基础研究与开发，激发人类的想象力与创造力，尤其是在娱乐领域。人工智能伦理是公司各项工作的核心，公司致力于开发能够在影像、计算机视觉和游戏等关键领域安全使用的新工具与新技术。

4 说话者简介

Orestis Papakyriakopoulos 是索尼人工智能的研究科学家，研究语音与自然语言处理技术的伦理问题。他的研究揭示政治议题，并为构建公正、包容且具有参与性的社会—算法生态系统提供理念、框架和实用解决方案。Orestis 的研究成果发表于 WWW、ICWSM、FAccT、AIES、EAMMO 等会议，他也经常在这些会议担任主席或程序委员会成员。他的研究还被《金融时报》和《华盛顿邮报》等众多新闻媒体引用。

参考文献

[1] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。2020。wav2vec 2.0：一种用于语音表征自监督学习的框架。神经信息处理系统进展 33 (2020)，12449–12460。[2] Alan W Black。2019。Cmu wilderness 多语言语音数据集。载于*2019* 年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，5971–5975。[3] Guoguo Chen、Shuzhou Chai、Guanbo Wang、Jiayu Du、Wei-Qiang Zhang、Chao Weng、Dan Su、Daniel Povey、Jan Trmal、Junbo Zhang 等。2021。Gigaspeech：一个不断扩展的多领域自动语音识别语料库，包含 10,000 小时的转录音频。*arXiv* 预印本 *arXiv:2106.06909* (2021)。[4] Siyuan Feng、Olya Kudina、Bence Mark Halpern 和 Odette Scharenborg。2021。量化自动语音识别中的偏差。*arXiv* 预印本 *arXiv:2103.15122* (2021)。[5] Allison Koenecke、Andrew Nam、Emily Lake、Joe Nudell、Minnie Quartey、Zion Mengesha、Connor Toups、John R Rickford、Dan Jurafsky 和 Sharad Goel。2020。自动语音识别中的种族差异。《美国国家科学院院刊》117，14 (2020)，7684–7689。[6] Nina Markl。2022。语言变异与算法偏差：理解英国英语自动语音识别中的算法偏差。载于*2022* 年 *ACM* 公平、问责与透明度会议。521–534。[7] Joshua L Martin。2021。口语语料数据、自动语音识别以及对非裔美国人语言的偏见：以习惯性“Be”为例。载于 *2021* 年 *ACM* 公平、问责与透明度会议论文集。284–284。[8] Kenny Peng、Arunesh Mathur 和 Arvind Narayanan。2021。减轻数据集造成的危害需要妥善管理：来自 1000 篇论文的经验教训。神经信息处理系统进展 (2021)。[9] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。2022。通过大规模弱监督实现稳健的语音识别。技术报告。技术报告，OpenAI。[10] Changhan Wang、Morgane Riviere、Ann Lee、Anne Wu、Chaitanya Talnikar、Daniel Haziza、Mary Williamson、Juan Pino 和 Emmanuel Dupoux。2021。Vox-populi：用于表征学习的大规模多语言语音语料库，

半监督学习与解读。*arXiv* 预印本 *arXiv:2101.00390* (2021)。

[11] Gary Yeung 和 Abeer Alwan。2018 年。关于幼儿园年龄儿童自动语音识别的困难。*Interspeech 2018* (2018)。