文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

边缘计算与隐私保护在方言数据采集与处理中的应用研究

陈后松　王可越　李卓林　邹晓芳　袁平*
广西职业师范学院
DOI:10.32629/acair.v3i4.17903

[摘　要] 方言作为地域文化的重要载体,其快速消亡正引发对语言多样性保护的迫切关注。传统采集方法受限于效率低下、传输压力与隐私泄露风险,难以适应大规模、可持续的方言保护需求。为此,本研究提出一种融合边缘计算与隐私保护技术的协同处理架构。该架构通过“终端—边缘—云端”三级分层设计,将音频特征提取、数据清洗等预处理任务下沉至边缘节点,显著减轻了网络带宽与云端存储负担;并系统集成差分隐私、联邦学习及同态加密等技术,实现从数据采集、传输、存储到分析的全流程隐私防护。实验结果表明,该方案可大幅降低数据传输延迟与存储开销,同时有效遏制针对用户声纹等敏感信息的推理攻击,在保障方言关键特征可用性的前提下,显著提升了数据采集的安全性与效率。本研究为方言资源的数字化保存与合规利用提供了一种高效且安全的技术路径。
[关键词] 方言数据采集；边缘计算；隐私保护；分布式处理；语言资源保护
中图分类号：G623.58　**文献标识码**：A

Research on the Application of Edge Computing and Privacy Protection in Dialect Data Collection and Processing

Housong Chen　Keyue Wang　Zhuolin Li　Xiaofang Zou　Ping Yuan*
Guangxi Vocational Normal University

[Abstract] As an important carrier of regional culture, the rapid decline of dialects has sparked urgent concern for the protection of linguistic diversity. Traditional collection methods, constrained by low efficiency, transmission pressure, and privacy leakage risks, struggle to meet the demands of large-scale and sustainable dialect preservation. To address this, this study proposes a collaborative processing architecture integrating edge computing and privacy protection technologies. Through a three-tiered "terminal-edge-cloud" design, the architecture offloads preprocessing tasks such as audio feature extraction and data cleaning to edge nodes, significantly reducing network bandwidth and cloud storage burdens. It systematically incorporates differential privacy, federated learning, and homomorphic encryption to achieve end-to-end privacy protection across data collection, transmission, storage, and analysis. Experimental results demonstrate that this solution can substantially reduce data transmission latency and storage overhead while effectively mitigating inference attacks targeting sensitive user information like voiceprints. By ensuring the usability of dialects' key features, it significantly enhances data collection security and efficiency. This research provides an efficient and secure technical pathway for the digital preservation and compliant utilization of dialect resources.
[Key words] dialect data collection; edge computing; privacy protection; distributed processing; language resource preservation

引言

方言作为地域文化的“活化石”,承载着民族语言多样性的核心价值,也是语言演化与文化传承的重要载体。然而,在全球化和城市化快速推进的背景下,方言正面临严峻的消亡危机。据相关统计,我国境内分布有超过一百种汉语方言和数百种少数民族语言,其中不少方言使用人口已不足万人,部分少数民族语言甚至缺乏文字记录,仅依靠口耳相传,濒临消失。方言的消失不仅是语言资源的流失,更意味着地方文化记忆的断裂和语言研究数据的永久缺失。传统的方言数据采集方式主要包括人工田野调查和中心化平台上传两种模式。前者受限于人力、经费与时间,覆盖范围狭窄,采集效率低下；后者虽然借助移动互联网提升了采集规模,但面临着数据传输延迟大、隐私泄露风险

高、存储成本昂贵等问题。尤其在偏远方言区,网络基础设施薄弱,上传大规模音频数据极为困难[1]。与此同时,方言数据中包含大量个人敏感信息,如声纹特征、地理位置、身份信息等,一旦在传输或存储过程中遭到泄露,将严重侵害用户隐私,甚至引发信息安全事件。因此,如何构建一个既高效又安全的方言数据采集与处理体系,已成为语言保护和语言学研究中的一项紧迫课题。

本文立足于方言数据采集的实际需求与技术现状,提出一种融合边缘计算与多层次隐私保护的协同处理架构。该架构通过将计算任务下沉至靠近数据源的边缘节点,有效缓解网络传输压力,提升数据处理效率；并综合运用差分隐私、联邦学习、同态加密等隐私计算技术,实现从数据采集、传输、存储到分析的全流程隐私保护。本文从技术层面探讨了该架构的设计与实现,旨在为方言资源的数字化保护提供一种可行、可靠的技术路径。

1 核心技术基础与方言数据特性分析

1.1边缘计算：面向场景的分布式处理范式

边缘计算的核心在于将计算与存储资源部署于物理上接近数据源的网络边缘,以此构建一种去中心化的分布式处理架构。在方言数据采集的具体场景中,这一范式展现出其不可替代的优势。它通过在本地区域(如乡镇文化站点)设立具备中等算力的节点,使原始音频数据无需经历长距离、高延迟的网络传输至云端,即可在本地完成清洗、特征提取等密集型预处理任务。这从根本上缓解了偏远方言区普遍存在的网络带宽不足问题,同时显著降低了系统整体响应延迟。更重要的是,边缘节点具备一定的离线工作能力,在网络间歇性中断的情况下仍可持续进行本地处理与缓存,待连接恢复后再同步处理结果,从而保障了数据采集任务的连续性与鲁棒性[2]。因此,边缘计算并非对云计算的简单替代,而是构成了一个协同的、层次化的计算体系,为方言这种地域性强、数据量大的非结构化资源处理提供了理想的底层支撑。

1.2隐私计算：保障数据全生命周期的安全技术体系

鉴于方言数据蕴含用户声纹、地理位置等敏感信息,构建覆盖数据全生命周期的隐私保护体系至关重要。当前主流的隐私计算技术为这一目标提供了多元化的实现路径。差分隐私通过向查询结果或数据集注入经过数学严谨定义的随机噪声,确保单个用户的存在与否不会对输出结果产生显著影响,从而在数据发布与统计分析中实现个体不可区分性。联邦学习则采用“数据不动模型动”的协作模式,各方仅交换模型参数(如梯度更新)而非原始数据,共同训练一个全局模型,这尤其适用于跨区域、跨机构的方言模型联合构建,能在保护数据隐私的前提下汇聚知识。同态加密允许对密文数据进行特定形式的运算,并获得加密后的结果,解密后即相当于对明文进行了相同运算,从而实现了数据“可用不可见”,为云端安全处理敏感特征提供了密码学保障[3]。这些技术并非互斥,在实际架构中往往协同应用,形成互补的防御层次,共同应对采集、传输、存储与分析各环节可能面临的隐私威胁。

1.3方言数据的多维特性及其技术需求

方言数据作为一种特殊的语言资源,具有区别于标准普通话的结构化语音数据的鲜明特征,这对处理技术提出了针对性需求。首先,其呈现显著的非结构化特性,数据主体为包含大量口语化表达、副语言特征(如语气、停顿)和地方俚语的连续音频,缺乏规整的文本对应,这就要求处理系统必须具备强大的音频信号处理与自动语音识别能力,以从中抽取有效的语言学特征。其次,方言与地域文化紧密绑定,呈现出强烈的空间异质性,同一方言在不同片区可能在音韵、词汇、语法上存在系统性差异[4]。因此,有效的技术方案必须能够融合地理信息系统信息,或具备基于地域知识进行自适应分析的能力。最后,方言数据具有高度的隐私敏感性。一段录音不仅包含可用于生物识别的声纹特征,其背景环境音、交谈内容也可能间接泄露说话者的身份、社会关系及活动轨迹。这意味着隐私保护必须是多维度的,需同时对声学特征、转写文本及元数据实施脱敏或访问控制,超越传统的单一维度保护思路。这些特性共同决定了,一个适用于方言数据的技术框架必须是能够协同处理非结构化信息、集成地理上下文并进行细粒度隐私管控的复合型系统。

2 系统架构设计与隐私保护实现

2.1“终端-边缘-云端”三级协同架构设计

为系统性地解决方言数据采集中的效率与隐私难题,本研究设计并实现了一种“终端-边缘-云端”三级协同处理架构。该架构的核心思想在于依据数据处理的逻辑与安全需求,将任务进行合理的层次化分解与部署。在终端层,广泛分布的智能手机与专用录音设备承担最前沿的数据采集任务。为减轻后端压力并初步保护隐私,终端集成了轻量级AI模型,能够在本地完成音频质量检测、基础降噪以及对明显个人身份信息的实时模糊化处理,并严格遵循最小数据收集原则,仅获取模糊化的用户属性。处于中间层的边缘节点是本架构的关键,通常部署于县域数据中心或乡镇文化站。它们具备远强于终端、但弱于云中心的适中算力,负责接收来自多个终端的数据流,执行计算密集型的深度数据清洗、高精度声学与语言学特征提取,以及基于差分隐私的核心脱敏操作。这一设计使得原始音频数据无需离开本地区域,既极大缓解了网络传输压力,也显著降低了数据在广域网上泄露的风险。云端层则作为全局大脑,负责聚合来自众多边缘节点的、已经过脱敏和标准化处理的特征数据,利用联邦学习技术进行分布式模型训练,并构建宏观的方言知识图谱与资源库,实现数据的最终价值汇聚与知识发现[5]。

2.2覆盖数据全生命周期的隐私保护机制

本研究的隐私保护机制贯穿于数据从产生到销毁的每一个环节,形成一套完整的闭环防御体系。在数据采集与授权阶段,系统通过交互式界面提供清晰、可选的隐私协议,引入动态知情同意模式,允许用户根据自身意愿选择不同颗粒度的匿名化等级(如是否允许保留年龄段信息),并赋予其对已提交数据的便捷撤回权,从而将隐私控制的主动权交还给用户。在数据传输过程中,终端与边缘、边缘与云端之间的所有通信链路均采用强加

密协议(如TLS 1.3)进行保护,并对所有传输的数据对象进行去标识化处理,使用无法关联至真实用户的随机标识符,有效防范传输过程中的窃听与流量分析攻击。在数据存储层面,系统实施分级存储策略：包含原始声纹的音频数据仅在边缘节点进行极短期的临时缓存,完成特征提取后即安全擦除;经过脱敏处理的核心特征向量则加密存储于云端,其加密密钥采用分布式门限方案进行管理,确保任何单一实体都无法独立解密数据,极大提升了存储的安全性。在最终的数据分析与使用阶段,所有对数据集的访问均实行基于角色的细粒度权限控制,并对发布的任何统计信息(如词频、音素分布)强制施加差分隐私保护,通过注入经过数学校准的噪声,确保分析结果不会泄露任何个体参与者的具体信息,实现了数据可用性与隐私性的严格平衡。

2.3面向资源受限环境的轻量化与自适应优化

考虑到实际部署中边缘节点往往存在计算资源、存储空间和能源供给方面的限制,本研究对核心处理模块进行了针对性的轻量化设计与自适应优化。在特征提取模型方面,通过采用深度可分离卷积替代传统卷积、对模型权重进行低精度量化(如从FP32至INT8)、以及利用模型剪枝去除冗余参数等一系列模型压缩技术,在保证特征提取精度的前提下,将模型体积与计算开销降低了约60%-70%,使其能够流畅运行在树莓派等主流边缘计算设备上。同时,算法层面支持利用边缘设备的异构计算能力(如GPU、NPU)对处理流程进行加速。更为关键的是,系统引入了动态隐私预算分配策略。系统能够根据数据内容本身(如是否检测到高辨识度声纹)、数据主体属性(如是否为未成年人)以及数据来源的地理敏感性(如是否位于特殊文化区域)等多个维度,自动化地、动态地调整差分隐私技术中的核心参数(如隐私预算ε)。这种自适应机制避免了“一刀切”式保护带来的效用损耗,能够在满足差异化隐私保护刚性要求的同时,最大化数据特征的可用性,实现了隐私保护强度与数据科学价值之间的精细化、智能化权衡。

3 总结

本研究针对方言数据采集中长期存在的效率瓶颈与隐私泄露风险,提出了一种深度融合边缘计算与多层次隐私保护技术的协同处理框架。该框架通过构建“终端-边缘-云端”三级分层架构,将计算密集型预处理任务有效下沉至网络边缘,实现了数据在源头附近的本地化处理与隐私增强,显著缓解了偏远地区的网络传输压力与云端存储负担。同时,通过系统整合差分隐私、联邦学习与同态加密等技术,构建了贯穿数据采集、传输、存储与分析全生命周期的闭环隐私防护体系,在确保方言关键语言学特征得以保留的前提下,有效遏制了对用户声纹、位置等敏感信息的推断与还原攻击,为方言资源的规模化、合规化采集奠定了坚实的安全基础。

本项目由国家级大学生创新创业训练计划项目资助,项目名称：广西职业师范学院2025年大学生创新创业训练计划项目《方的言-基于人工智能的地方方言传承与双向翻译平台》,项目级别：国家级,项目类别：一般项目,项目编号：202514684013X。

[参考文献]

[1]久西草,更太加,金弟,等.多模态技术在藏语安多方言生理语音研究中的应用进展[J].青海科技,2025,32(04):180-188.

[2]吕晓东.面向边缘智能的AI服务管理系统的研究与实现[D].北京邮电大学,2024.

[3]曹达钦,刘世界,陈昕.数字化时代语料数据伦理研究：概念、问题、原则与路径[J].外语电化教学,2025,(05):23-29+104.

[4]翁毅,王璐,孔江平.双方言语音及嗓音特征分析研究——以粤—普人群为例[J].中国语音学报,2024,(02):65-74.

[5]杜逸超.面向低资源场景的端到端语音和文本翻译方法研究[D].中国科学技术大学,2025.

作者简介：

陈后松(2002--),男,汉族,广西桂平人,本科,单位：广西职业师范学院,研究方向：物联网工程。

王可越(2002--),女,汉族,广西桂平人,本科,单位：广西职业师范学院,研究方向：物联网工程。

李卓林(2001--),男,汉族,广西博白人,本科,单位：广西职业师范学院,研究方向：物联网工程。

邹晓芳(2005--),女,汉族,广西钦州人,本科,单位：广西职业师范学院,研究方向：信息管理与信息系统。

*通讯作者：

袁平(1979--),男,壮族,广西南宁人,硕士研究生,单位：广西职业师范学院,研究方向：信息安全。