arXiv:2601.02209v1 [cs.CL] 5 Jan 2026

ARCADE：用于细粒度阿拉伯语方言标注的城市规模语料库

Omer Nacar[1], Serry Sibaee[2],Adel Ammar[2*], Yasser Alhabashi[2],Nadia Samer Sibai[2], Yara Farouk Ahmed[2], Ahmed Saud Alqusaiyer[2],Sulieman Mahmoud AlMahmoud[2],Abdulrhman Mamdoh Mukhaniq[2], Lubaba Raed[2],Sulaiman Mohammed Alatwah[2],Waad Nasser Alqahtani[2],Yousif Abdulmajeed Alnasser[2],Mohamed Aziz Khadraoui[3],Wadii Boulila[2]

[1]Tuwaiq Academy，沙特阿拉伯利雅得 13415[2]机器人与物联网实验室，苏丹王子大学，沙特阿拉伯利雅得[3] 突尼斯市高等通信学院（SUP'COM），突尼斯 Ariana 2083[*]通讯作者：aaammar@psu.edu.sa

摘要

阿拉伯语具有丰富多样的区域方言，这些方言在语音学和词汇方面存在显著差异，反映了其使用者在地理和文化上的多样性。尽管已有许多多方言数据集，但将语音映射到细粒度方言来源（如城市）这一方向仍未得到充分探索。我们推出 ARCADE（阿拉伯语音频方言评测广播语料库），这是首个专为城市级方言粒度设计的阿拉伯语语音数据集。该语料库由从阿拉伯世界各地的流媒体服务收集的阿拉伯语广播语音构成。我们的数据处理流程从经过核实的广播流中截取 30 秒片段，涵盖现代标准阿拉伯语（MSA）和多种方言语音。为确保可靠性，每个片段均由一至三名阿拉伯语母语审校员进行标注，并赋予丰富的元数据，包括情绪、语音类型、方言类别，以及用于方言识别任务的有效性标记。最终语料库包含 6,907 条标注和 3,790 个独特音频片段，覆盖 19 个国家的 58 座城市。这些细粒度标注有助于稳健的多任务学习，并可作为城市级方言标注的基准。我们详细介绍了数据收集方法，评估了音频质量，并全面分析了标签分布。

关键词 阿拉伯语方言识别 ·方言分类 · 广播数据收集 · 多方言语音

1 引言

阿拉伯语在其众多方言中呈现出显著差异，这些方言不仅在词汇和形态上有所不同，在音系上也各不相同。古典阿拉伯语和现代标准阿拉伯语（MSA）构成了通用的正式书面标准，并见于部分正式媒体；然而，所有阿拉伯国家的日常语音都使用方言语。阿拉伯方言语（DA）在发音、词汇和语法方面均不同于现代标准阿拉伯语和古典阿拉伯语，反映了说话者的地理来源 [1]。关于综合阿拉伯语方言数据集（IADD）的先前研究指出，阿拉伯方言语与古典阿拉伯语和现代标准阿拉伯语存在显著差异，因为它反映了说话者的地理位置 [2]。因此，我们构建了一个包含细粒度城市级录音的语料库，以支持有关方言变异、社会语言学和文化对齐语音模型的研究 [3]。

阿拉伯语方言识别的先前研究通常在国家或大型地区尺度上进行分类，其语料库和评估也遵循这一做法。这种设置能够提供较高的基线性能，并支持跨系统可比性。然而，它对城市级标注的支持有限，阻碍了对城市与农村言语的分析，也限制了能够从更细粒度方言信号中获益的应用开发。

Nuanced Arabic Dialect Identification（NADI）共享任务提供基准数据，用于将话语归类为国家级方言标签，并推动了方言识别领域最先进水平的发展 [4]。例如，2024年

这一届引入了多标签和多层级方言识别任务，但表现仅属中等 [4]。最近，NADI 2025 将重点从文本转向语音处理，为口语方言识别、自动语音识别（ASR）和添加变音符号建立了统一基准 [5]。尽管表现最佳的识别系统准确率达到 79.8%，但结果凸显了处理多方言阿拉伯语口语仍面临的挑战。在音频层面，ADI17 数据集收集了约 3,000 小时的 YouTube 语音，涵盖 17 个阿拉伯国家，用于方言识别 [6]。这些数据集提供的是粗略的地理标签（通常为国家或地区级别），仅限于高层级方言类别，且几乎不包含有关国家内部差异的信息。

相比之下，本文提出了一个由短广播片段构成的语料库，这些片段带有城市级标签和详细元数据，支持细粒度方言建模与评估。我们的贡献主要有三点：

- 我们设计并实现了一套数据处理流程，从众多阿拉伯城市采集广播流，筛选独白片段，并为片段标注情绪、语音类型、方言类别和质量标签。
- 我们分析了所得语料库，展示了情绪、语音类型、保留/跳过决策、方言类别和置信度标签的分布。该数据集涵盖19个阿拉伯国家的58座城市，每座城市至少包含10条录音，提供了广泛的地理覆盖。
- 我们概述了细粒度建模的实际应用场景和关键注意事项。基准测试和建模留待未来工作开展。

通过发布数据以及透明的数据收集与标注协议，我们旨在推动基于音频的细粒度方言归属判定研究。该语料库支持城市级标注、方言来源建模、跨通道和领域偏移的鲁棒性研究，以及向电话通信和在线视频等相关媒体的迁移。可重用的协议支持未来扩展、社区贡献，并为公平的建模比较提供可复现基准。

2 相关研究

本节回顾阿拉伯语方言识别研究的发展历程，涵盖文本和声学两个视角。早期研究主要依赖基于文本的语料库，从区域级数据集到后来的多标签基准，揭示了阿拉伯方言语与现代标准阿拉伯语之间系统性的差异。近来，研究重点转向口语数据，大规模广播和网络来源的语音语料库推动了声学建模的发展。尽管取得了这些进展，现有资源仍受限于粗粒度的方言标签和参差不齐的标注质量，因此亟需构建方言标注更细致、可靠的数据集。

2.1 方言识别的文本资源

阿拉伯语方言识别研究历来由文本资源推动。早期语料库，如 PADIC[7], SHAMI [8, 1], 和 TSAC [9]，侧重于基于文本的方言分类。此后，这些语料库被整合进更大型的基准，例如综合阿拉伯语方言数据集（IADD） [2],。该数据集合并了五个公开数据集的子集，涵盖黎凡特方言、突尼斯方言、埃及方言、马格里布方言、伊拉克方言和海湾方言。这项研究表明，阿拉伯方言语（DA）与现代标准阿拉伯语（MSA）及古典阿拉伯语（CA）存在系统性差异，体现为各具特色的区域变体，而这些差异对方言识别至关重要。近期研究通过开发通用文本嵌入，进一步推动了阿拉伯语自然语言处理，以更好地捕捉语义和方言细微差异 [10]；然而，将这些文本领域的研究成果应用于声学领域仍是一项独特的挑战。

共享任务在推动文本方言识别方面也发挥了关键作用。Nuanced Arabic Dialect Identification（NADI）系列最初在文本层面开展，为推文提供国家级标签。例如，2024年版引入了多标签国家识别和方言强度预测任务 [4]。尽管取得了这些进展，基于文本的识别仍不同于口语方言中声学变异所带来的挑战。

2.2 口语阿拉伯语语料库与基准

稳健的自动语音识别（ASR）和口语方言识别系统的发展，在很大程度上依赖于大规模广播语料库和网络爬取语料库。早期工作，例如 CALLHOME 埃及阿拉伯语语料库 [11], ，为电话会话语音提供了基础资源。

近来，多体裁广播（MGB）挑战赛系列为该领域建立了重要基准。MGB-1 [12] 聚焦英语，而 MGB-2 [13] 则推出了一项规模庞大的阿拉伯语资源，包含超过 1,200 小时的半岛电视台广播节目，涵盖现代标准阿拉伯语和四大方言地区（埃及方言、海湾方言、黎凡特方言和北非方言），

而 MGB-3 [14] 则将重点转向“自然环境中的”方言语音，利用多体裁 YouTube 视频，主要用于埃及阿拉伯语自动语音识别。
为了解决相对细粒度的识别问题，ADI17语料库 [6] 收集了来自17个国家、约3,000小时的YouTube音频。随后，该数据集被用于以国家级方言识别为目标的MGB-5挑战赛 [15], 。类似地，卡塔尔计算研究所阿拉伯语语音识别（QASR）数据集 [16] 提供了2,000小时的广播语音，并包含语言学分段和命名实体标注。

表1对这些主要资源进行了比较总结。尽管近年来取得了进展，大多数现有语料库提供的方言标签仍较为粗略，仅标注到国家或区域级别（如埃及方言、黎凡特方言）。此外，标签可靠性也存在显著差异。MGB-5 [15] 和 ADI17 [6] 等大规模数据集依赖弱监督，即根据频道元数据推断方言标签，这不可避免地会给训练数据[15]带来噪声。类似地，QASR依靠轻量监督进行转录，并通过自动分类开展方言分析，只有一小部分数据经过人工标注，用于国家级评估 [16]。
尽管我们的数据集在时长或录音数量方面并不是最大的，但其主要增值在于粒度、广度和标签质量的结合。我们的工作通过为数量最多的阿拉伯国家（19个）提供经过细粒度划分的次区域方言标签而形成差异；这些标签由11名经过培训的阿拉伯语母语标注员进行严格验证，每个音频片段包含一至三条标注。这一全面的标注流程记录了详细的元数据，包括说话人情绪、语音类型（例如单个/多个说话人、《古兰经》诵读、音乐）、内容有效性（保留/跳过）、方言类别（现代标准阿拉伯语、方言、混合）以及标注员置信度，从而弥补了地理定位精确且可靠的方言识别资源方面的关键空白。

表1：ARCADE与主要现有阿拉伯语语音资源的比较。请注意依赖噪声/弱标签的资源与具有经验证标注的资源之间的区别。

资源	规模（片段数/时长）	Type	重点	粒度	标注	国家
ADI17 [6]	~1百万个片段（3,000 小时）	17 种方言 现代标准阿拉伯语 +4 个地区	YouTube（细粒度）	国家级	弱（基于频道）	17
MGB-2 [13]	~550k 个片段（1,200 小时）		广播（半岛电视台）	区域级	弱监督	–
MGB-3 [14]	13,825 个片段（16 小时）	埃及阿拉伯语 +5 方言	YouTube（多体裁）	区域级	人工验证	5
MGB-5 [15]	~1.06M 片段（3,000 小时）	17 种方言 现代标准阿拉伯语 +4 地区	YouTube（细粒度）	国家级	弱（基于频道）	17
QASR [16]	160万条片段（2,000小时）		广播（半岛电视台）	区域级	弱监督	–
CALLHOME [11]	120通电话（~60 小时）	埃及方言	电话语音	特定方言	人工验证	1
ARCADE（我们的）	3,790 个片段（31.6 小时）	现代标准阿拉伯语 + 方言	广播流	**城市级**	**人工（多项标注）**	**19**

3 数据集收集与标注

3.1 录音流程

我们开发了一套流程，可自动发现并录制阿拉伯世界各地的广播流（图 1）。流媒体来源包括 Radio Garden[1], World Radio Map[2], 以及其他公共电台目录。我们会验证每个流，确认音频可访问且质量足以用于分析。验证通过后，系统会以随机时间间隔录制 30 秒的片段。我们特意选择这一时长，以兼顾稳健性与同质性；先前研究表明，30 秒的片段在阿拉伯语方言识别任务中表现出色（准确率超过 81%） [17], ，同时也更有可能确保每段录音仅包含一位说话者，且语音风格一致。我们以独白式语音为目标，并确保录制时间段丰富多样，以捕捉不同的节目和主持人（新闻、综艺节目、文化片段及类似形式）。片段以结构化文件名存储，其中包含城市和时间戳信息（<city>_<timestamp>.mp3）。每个片段记录的元数据包括电台名称、城市、国家、开始时间和流媒体网址。

为扩大覆盖范围，我们规定每个城市至少保留 10 条录音，并力求每个国家至少抽取 5 个城市；然而，由于一些国家缺少 5 个来自不同城市且在互联网上公开广播的可识别广播电台，这一目标无法始终实现。城市的选择依据人口规模、媒体影响力和地理分布。数据集由 11 名经过培训的阿拉伯语母语标注员完成标注，他们识别说话者情绪（例如中性、快乐或愤怒），标注语音类型（普通单说话者、多个说话人、无语音/音乐或《古兰经》诵读），判断片段应保留还是跳过，并确定方言类别（现代标准阿拉伯语、方言、混合或不适用）；此外，他们还提供置信度（确定、不确定或不知道）以核实分类结果，从而确保将在本地广播电台出现的非本地方言说话者标记出来。

[1]https://radio.garden/[2]
https://worldradiomap.com/

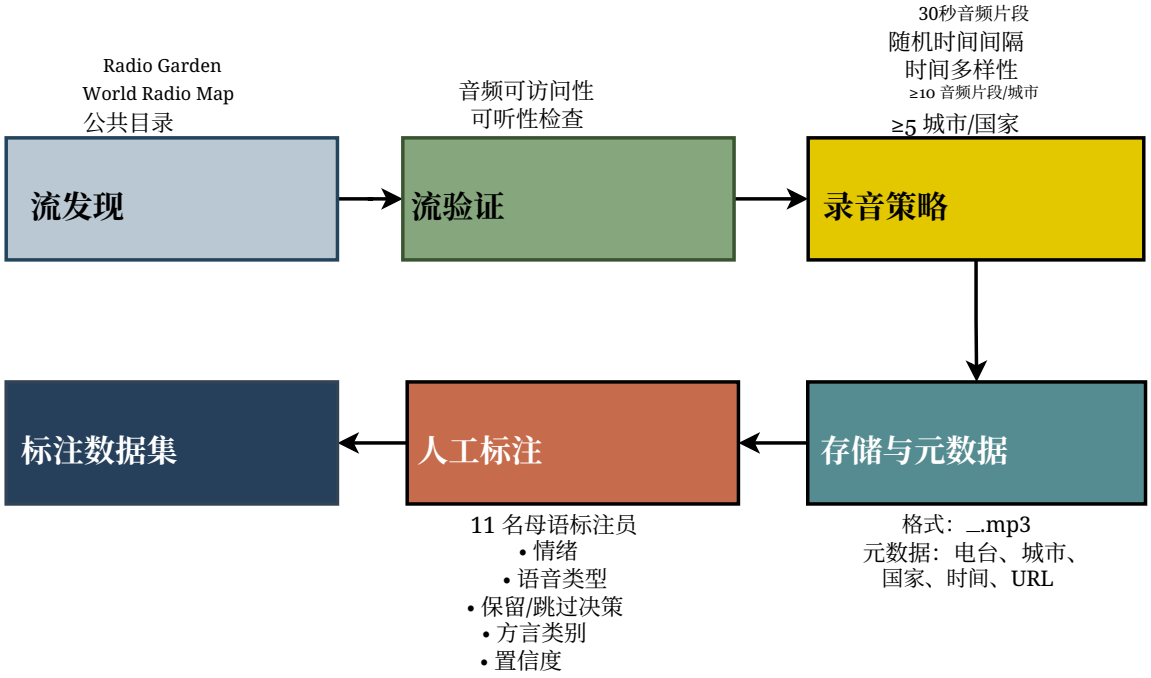

图1：构建 ARCADE 数据集的主要流程。

为支持高效标注，我们开发了一个基于 Gradio 框架构建的定制网页界面（图 2）。该界面支持标注员播放音频片段，同时通过标准化下拉菜单为情绪、语音类型、方言类别和置信度分配标签。界面集成了用于顺序标注或跳过的导航控件，并配备实时统计仪表板，以监控个人和整体进度，确保标注团队之间的数据一致性。

3.2 数据概览

当前版本包含 6,907 条标注和 3,790 个独立音频片段。其中，65.7% 经人工标记为“保留”，用于方言识别任务；其余 34.3% 则因《古兰经》诵读、音乐或串音占主导而在数据集中标记为“跳过”。标注后的分析显示，常规单人语音占片段总数的 37.3%，而 35.4% 的片段包含多个说话人；此外，22.7% 包含音乐或非语音音频，4.6% 被识别为《古兰经》诵读（图3）。后者被排除在外，因为宗教诵读遵循古典阿拉伯语严格且标准化的音系 [18],，这会抑制方言识别所需的音系组合变异。此外，广播节目经常播放知名的跨国诵经者，而非当地说话者，因此说话者的来源地与电台的地理位置并不对应。

表2介绍了 ARCADE 数据集中使用的标注方案。每个音频片段均以 MP3 格式存储，并记录相应的文件名 和时长（以秒为单位）。地理元数据记录了音频的国家 和城市来源信息，从而支持细粒度的城市级方言分析，使 ARCADE 有别于现有资源。核心语言标注用于说明每个片段包含的是现代标准阿拉伯语还是方言语音（现代标准阿拉伯语或方言），而音频类型 字段则将片段归类为包含普通说话者、多个说话者或无语音。情绪 字段记录说话者的情绪基调。通过保留或跳过字段筛除不合适的样本，并记录标注员自我报告的置信度，以实施质量控制。通过标注员 身份信息和时间戳字段确保完整的可追溯性，从而支持标注流程的可复现性。

如图4所示（由标注结果得出），方言语音占片段总数的41%，现代标准阿拉伯语占21%，混合语音占18%，不适用的占20%（音乐或背景噪声盖过了说话者的声音）。保留现代标准阿拉伯语片段，是因为即使说话者使用规范化的现代标准阿拉伯语，其地域来源通常仍可通过口音线索推断。事实上，音系干扰和韵律特征保留了地域身份，由此形成现代标准阿拉伯语中独特的“地域标准变体” [17, 19, 20, 21]。

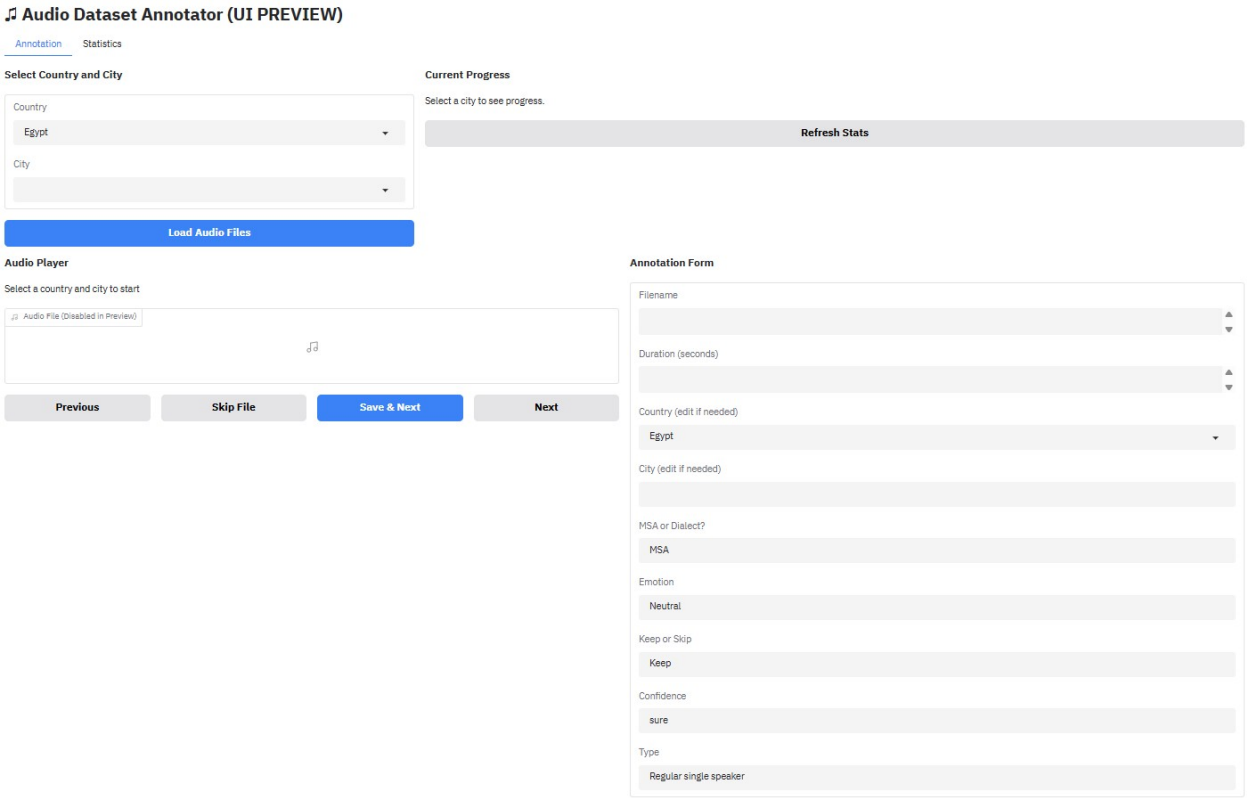

图 2：界面截图 母语者用于标注数据的定制 Gradio 标注界面截图 et.

表 2：数据集标注字段说明 s

字段	说明
音频	MP3 格式的音频文件。
文件名	音频文件的名称。
时长	音频片段的长度（秒）。
国家	说话者或录音的来源国家。
City	说话者或录音所在的城市。
现代标准阿拉伯语还是方言?	现代标准阿拉伯语或某种特定方言。
情绪	音频中表达的情绪语气（参见图5）。
保留或跳过	在方言识别任务中予以保留或舍弃。
置信度	标注员对标注结果的置信度（确定、不确定, 或完全不清楚）。
音频类型	语音/音频的类别或类型（普通说话人、多位说话者、《古兰经》诵读或音乐）。
标注员	标注该样本的标注员姓名。
时间戳	标注完成的时间。

图5总结了标注情绪的分布情况。中性情绪占主导地位，占标注总数的87.8%；而兴奋、快乐、悲伤、挫折、愤怒、惊讶和恐惧等其他情绪出现的频率要低得多。该数据集的不平衡情况将为任何下游情绪识别模型的采样策略提供参考。

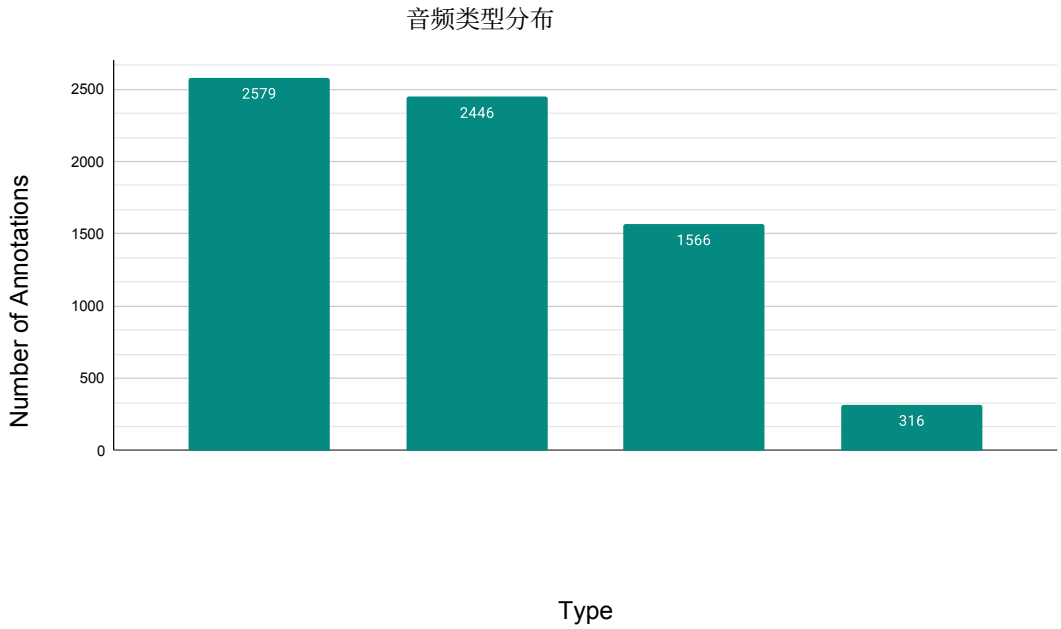

图3：标注语料库中各类音频的分布。该数据集主要由常规单说话者录音和多说话人片段组成，另有少量包含音乐/无语音或《古兰经》诵读的内容。

表3：ARCADE 数据集涵盖的国家和城市。

国家	城市
阿尔及利亚	阿尔及尔、安纳巴、巴特纳、布利达、比斯克拉、君士坦丁、杰勒法、奥兰、塞提夫
巴林	麦纳麦
埃及	亚历山大、开罗
Iraq	巴格达、巴士拉
约旦	安曼 ，伊尔比德
科威特	科威特城
黎巴嫩	贝鲁特，谢赫塔巴
利比亚	班加西、的黎波里
摩洛哥	卡萨布兰卡、非斯、马拉喀什、拉巴特、丹吉尔
Oman	马斯喀特，全国范围
巴勒斯坦	希伯伦，耶路撒冷，纳布卢斯，Gaza, 拉姆安拉
卡塔尔	Doha
KSA	吉达、麦加、利雅得
索马里	Daljir、摩加迪沙、谢贝利、索马里人
苏丹	欧拜伊德，喀土穆 ，恩图曼 ，苏丹港 ，瓦德迈达尼
叙利亚	阿勒颇、大马士革
突尼斯	突尼斯市
UAE	阿布扎比、阿治曼、迪拜、富查伊拉、沙迦
也门	Aden, 荷台达 ，萨那，Taiz

3.3 地理覆盖范围

我们从横跨非洲、黎凡特、伊拉克和海湾地区的58座城市**收集了音频片段**（所有城市均列于表3）。图6展示了所有保留的广播电台的地理分布。录音分布并不均衡，阿尔及利亚和巴林的片段数量尤为集中，分别贡献了1,509段和1,014段，其次是苏丹和卡塔尔（图7）。未来的工作或许需要通过有针对性的数据

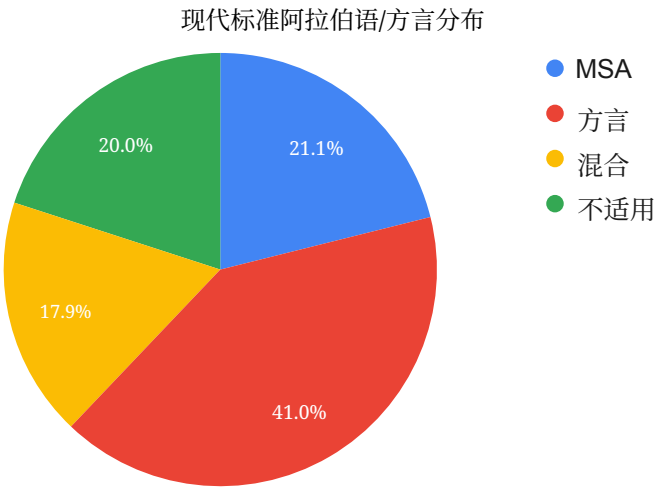

图4：所收集的标注片段中现代标准阿拉伯语/方言的分布情况。方言语音占据数据的最大部分，其次是现代标准阿拉伯语，然后是混合类别和不适用类别。

收集，或在模型训练期间使用加权损失函数来解决这种不均衡。我们将所有实验基准测试和建模工作留待未来研究，本文则聚焦于描述该语料库。

3.4 技术验证

为确保 ARCADE语料库的可靠性和实用性，我们进行了全面的技术验证，涵盖标注质量和声学信号特征两个方面。该验证涉及两个相辅相成的维度：一是人工标注的一致性和可靠性；二是录制音频的声学质量，通过信号级测量进行表征。这些分析共同确立了该语料库作为细粒度方言识别研究的可靠资源的地位，并向潜在用户说明可能影响建模决策的数据质量考量。

3.4.1 标注质量

每个音频片段均由受过培训的阿拉伯语母语者独立标注。我们最初计划每个片段由三名标注员标注，以确保高度可靠；然而，由于时间限制，部分审核人员未能完成分配给他们的批次，因此实际覆盖率有所不同。如图8所示，14.4%的语料库达到了三名审核人员的完整目标，53.4%的语料库由两名审核人员验证，为大多数数据提供了稳固的共识基础。其余32.2%的数据仅经过一轮标注。标注员标注了情绪、语音类型、质量判定（保留/跳过）、方言类别（现代标准阿拉伯语、方言、混合、不适用）和置信度（确定、不确定、不知道）。较高的置信度表明标注可靠：91.9%为“确定”，7.1%为“不确定”，仅1.0%为“不知道”（图9），从而有效标记出含糊案例和潜在的非本地方言说话者。在6,907个已标注片段中，经过人工审核后，有4,539个（65.7%）被保留用于建模；被排除的片段主要包含《古兰经》诵读、音乐或串音。

表4报告了各标注类别的标注者间一致性。保留/跳过决策的原始一致率最高（91.76%），Cohen κ系数为0.507，表明一致性中等；这说明尽管卡帕系数容易受到类别不平衡的影响，纳入标准总体上仍较为一致。语音类型类别的原始一致率也较高（87.71%），卡帕系数同样处于中等水平（0.586），反映出区分单人和多人语音片段时偶尔存在歧义，尤其是在包含简短插话的脱口秀形式中。

现代标准阿拉伯语与方言的区分和情绪标签表现出较高的百分比一致性（分别为 83.16% 和 90.53%），但卡帕系数较低（0.310 和 0.179）。这种表面上的差异源于众所周知的事实：Cohen κ系数对类别不平衡十分敏感 [22]：当某一类别占主导地位时（例如，中性情绪占 98%，或方言语音占多数），即使少数类别上只有中等程度的不一致，也会大幅拉低卡帕系数，而原始一致性仍然较高。就情绪而言，中性标签占压倒性多数，意味着标注员很少遇到非中性案例，因此在少数情绪类别上达成一致或产生分歧的机会都有限。对于

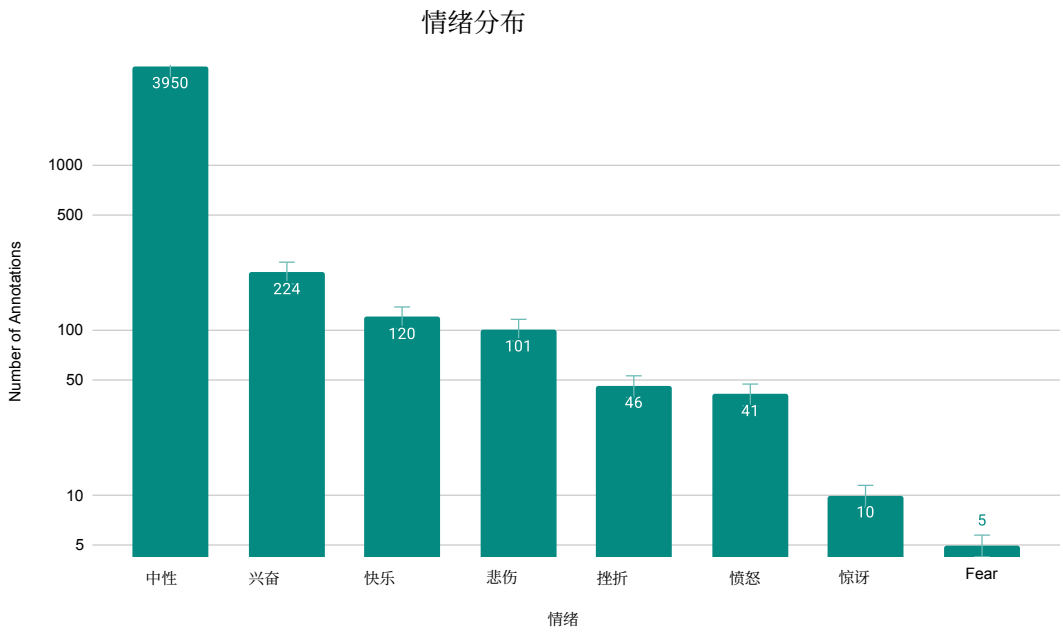

图5：ARCADE 数据集中的情绪分布（对数刻度）。情绪标注主要应用于保留的片段，因为跳过的音频片段（通常为《古兰经》诵读或音乐）不需要进行情绪标注。绝大多数标注为中性情绪，其他情绪则较为少见。请注意，42 段保留的录音缺少情绪标注，因此未纳入此图。

现代标准阿拉伯语与方言类别的标注难度则更高：当说话者进行语码转换或说出带有地区口音的现代标准阿拉伯语时，区分现代标准阿拉伯语和方言本身就很困难；这也是阿拉伯语语音语料库中标注者差异的一个公认来源。尽管卡帕系数中等，但较高的原始一致性率和标注员置信度表明，这些标签足以用于方言识别模型的训练与评估。

表4：标注者间一致性指标

标注类别	百分比一致性	Cohen κ系数
现代标准阿拉伯语还是方言?	83.16%	0.310
情绪	90.53%	0.179
保留或跳过	91.76%	0.507
Type	87.71%	0.586

3.4.2 音频质量

为描述所收集录音的声学属性，我们为每个片段计算了四项信号级指标：信噪比（SNR）、静音比例、动态范围和频谱质心。图10展示了这些指标在完整数据集中的分布。

信噪比分布大致呈正态形状，中心值为 13.24 dB，数值范围约为−10 dB 至 50 dB 以上。这一范围反映了广播节目的异质性，其中包括录音室质量的语音片段、现场录音，以及音频质量较差的听众热线节目。静音比例定义为低于语音活动阈值的帧所占百分比，其分布呈右偏态，平均值为 15.98%，表明大多数片段以活跃语音为主。动态范围平均为 15.83 dB，数值集中在 10–20 dB 之间，这与广播音频在传输过程中经过动态范围压缩的情况一致。频谱质心是衡量音频信号“明亮度”的指标，其中心值约为 1602 赫兹，这是以语音为主的内容的典型特征。

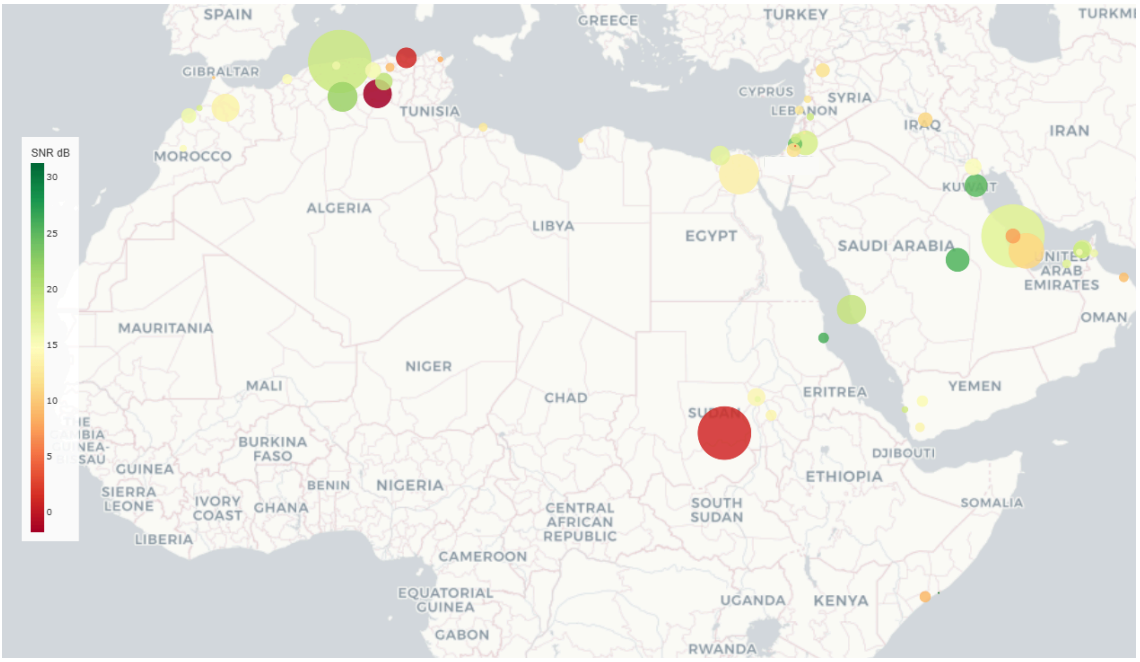

图6：数据集在阿拉伯世界的空间分布。标记表示城市位置；圆圈大小与收集到的录音量成正比，颜色表示各城市录音的平均信噪比。可通过 https://riotu-lab.github.io/arabic-cities-map/ 查看交互式可视化。

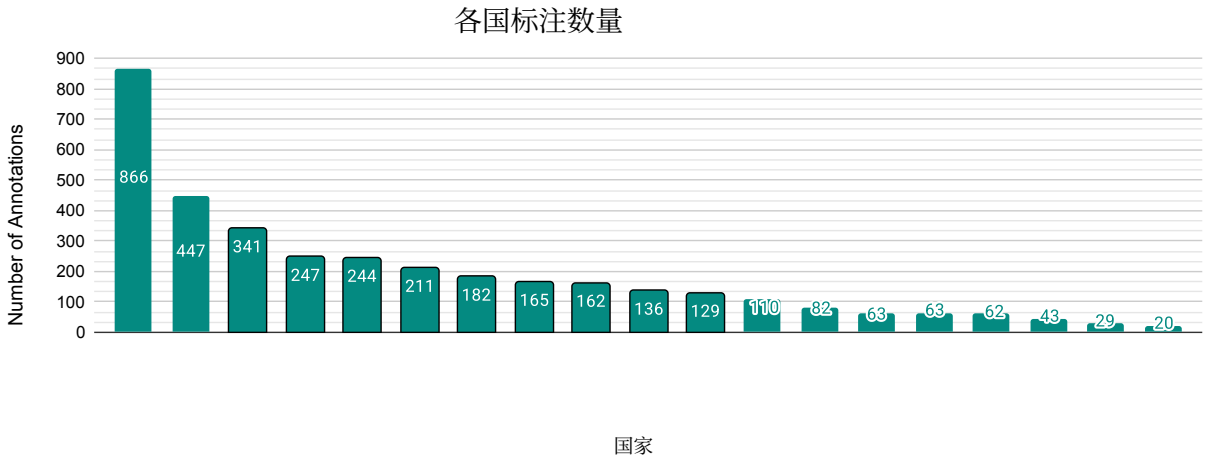

图7：各国标注录音的分布。阿尔及利亚和巴林的数据量最高（分别为 21.8% 和 14.7%），而阿曼和突尼斯的数据量最低（分别为 0.8% 和 0.6%）。

图11比较了保留（“保留”）与排除（“跳过”）片段的音频质量分布。值得注意的是，标注流程并未明确考虑音频质量；排除片段仅依据内容类型（《古兰经》诵读、音乐或串音），目的是使语料库专注于方言语音。尽管如此，作为基于内容进行过滤的副产品，保留的片段呈现出较好的声学属性。保留片段的平均信噪比更高（15.25 dB，而跳过片段为 9.36 dB），这可能是因为音乐和《古兰经》诵读片段通常包含背景伴奏或混响，从而降低信噪比估算值。保留片段的静音比例更高（19.30%，而跳过片段为 9.56%），反映了对话语音中自然停顿较多，而音乐通常连续不断。保留片段的动态范围也更高（16.94 dB，而跳过片段为 13.72 dB），频谱质心则更低（1535 赫兹，而跳过片段为 1731 赫兹），这与排除了音乐的情况一致，因为音乐的高频成分通常比语音更宽泛。这些附带的

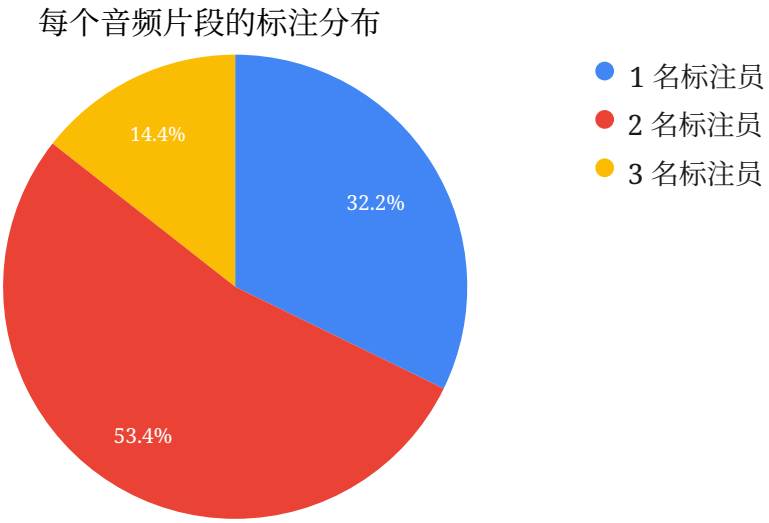

图8：每个音频片段的标注员数量分布。语料库中超过三分之二（67.8%）由多名标注员（2 或 3 名）验证，而 32.2% 仅经过一轮标注。

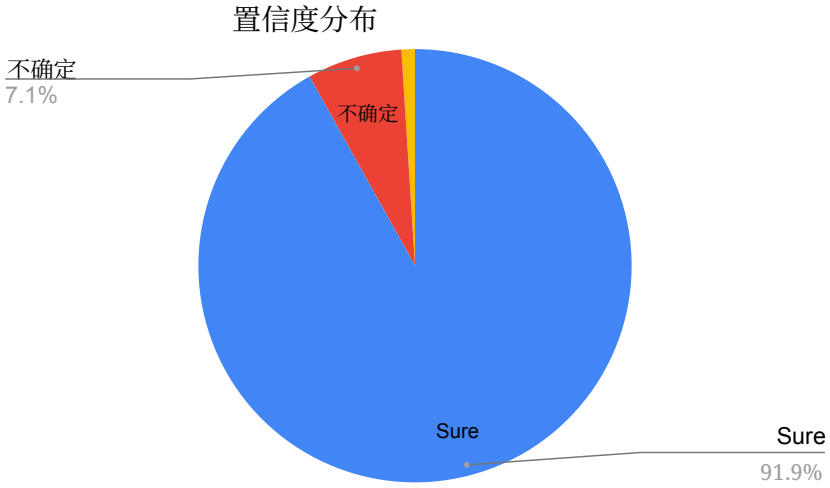

图9：标注员置信度等级。绝大多数标签均标记为“确定”，表明标注结果可靠。

质量提升是积极成果：以方言为重点的过滤筛选出声学质量高于平均水平的子集，这可能有利于下游建模。

音频质量的地理分布如图6所示，圆圈颜色表示各城市的平均信噪比，渐变范围从红色（信噪比较低，～0 dB）经黄色（～15 dB）到绿色（信噪比较高，～30 dB）。其中呈现出显著的区域模式：北非城市，尤其是摩洛哥和阿尔及利亚部分地区的城市，大多显示为绿色标记，表明录音质量较高（平均信噪比为 > 25 dB）。同样，利雅得和科威特也显示为绿色标记，反映出这些海湾国家首都拥有高质量的广播基础设施。相比之下，苏丹的一些城市和黎凡特部分地区显示为红色或橙色标记，表明信号质量较低，这可能与流媒体基础设施、电台设备或录音条件有关。海湾地区其他城市的音频质量参差不齐，阿联酋的电台通常达到中等至较高的信噪比数值。训练模型时应考虑这些音频质量的地理差异，因为它们可能会使声学劣化与方言特征相混淆。

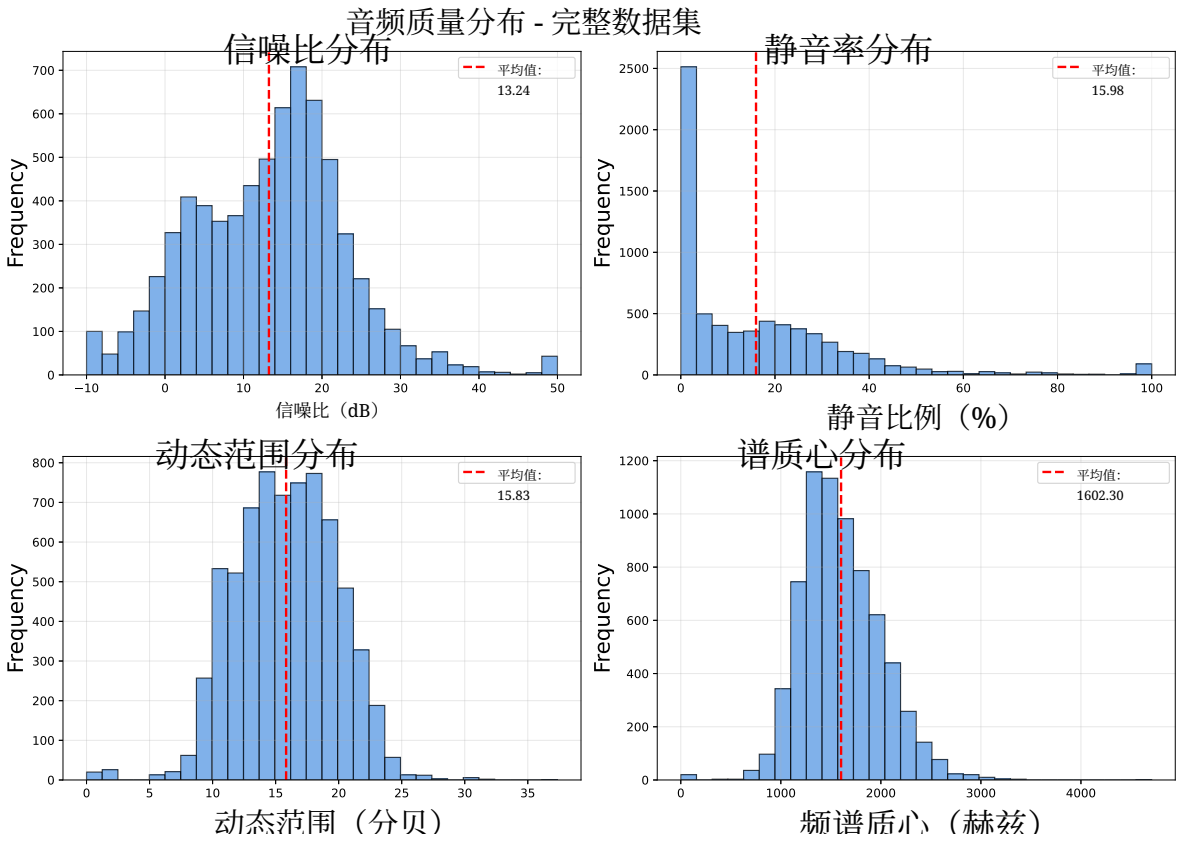

图10：完整 ARCADE 数据集中音频质量指标的分布。四个面板分别展示信噪比 (SNR)、静音比例、动态范围和频谱质心。红色虚线表示平均值。

4 讨论与未来方向

我们的广播语料库表明，可以通过公共流媒体服务收集地理覆盖丰富的阿拉伯语语音。对情绪、语音类型和方言类别进行标注，可支持多任务模型联合学习预测地理位置及其他属性。不过，仍有若干挑战有待解决。首先，数据在地理分布和情绪方面均不均衡，一些城市代表性过高，而积极情绪的占比不足。未来的采集工作应重点覆盖采样不足的地区（如农村地区），以及能够引发多样情绪的节目类型。其次，广播环境会引入噪声、音乐和串音；改进过滤或采用自适应去噪技术，有望提升信号质量。

展望未来，细粒度地理元数据与语言学标注的结合，为新的研究方向铺平了道路。例如，多任务学习框架可以联合预测方言类别、情绪和次区域来源，从而将语音线索与其他潜在因素区分开来。此外，将音频表示与空间标识符对齐的对比方法，也为未来研究提供了颇具前景的方向。我们强调，尽管这些建模工作超出了本研究的范围（本研究聚焦于数据收集和描述性分析），但我们希望发布这一语料库能够激发人们对全球音频地理定位和地理语言学更广泛的兴趣。

5 数据可用性

完整的 ARCADE 数据集可通过 Hugging Face Datasets 平台公开获取[3]。所有数据文件均依据开放研究许可证（知识共享署名 4.0 国际许可协议）发布，旨在用于非商业性的学术和研究用途。该存储库包含完整的音频语料库、标注文件（含情绪、语音类型、方言类别和质量标签）、详细文档，以及将每个片段关联至其来源城市和国家的元数据。此次结构化发布

[3]https:// huggingface.co/datasets/riotu-lab/ARCADE- full

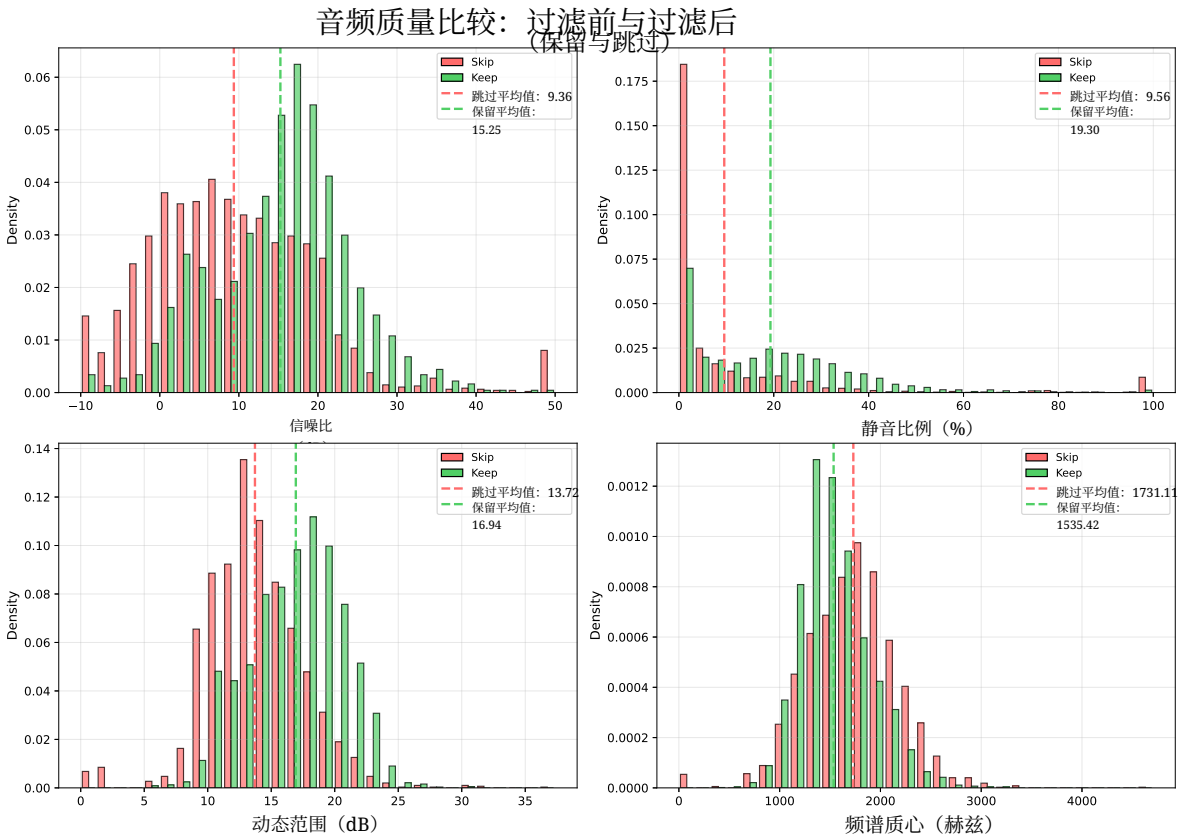

图11：保留（“保留”，绿色）与排除（“跳过”，红色）片段的音频质量指标比较。信噪比、静音比例、动态范围和频谱质心的系统性差异验证了人工过滤决策。

支持 FAIR数据原则（可发现、可访问、可互操作和可重用），并永久存档，以支持可复现研究。

6 结论

我们介绍了 ARCADE，这是一个新颖的阿拉伯语广播语音语料库，标注了城市级地理标签并附有丰富的元数据，填补了细粒度方言识别资源方面的关键空缺。与现有仅提供国家级或区域级标注的语料库不同，ARCADE 收录了来自19个阿拉伯国家58座城市的录音，使研究此前因数据限制而难以开展的次区域方言变异成为可能。

我们的采集流程利用 Radio Garden 和 WorldRadio Map 等平台公开提供的广播流，证明无需进行昂贵的现场录音，也能从现有广播基础设施中获取大规模、地域多样的语音数据。该流程确保每座城市至少有十段录音，共获得 3,790 个独特的 30 秒片段。每个片段均标注了情绪、语音类型、方言类别和标注员置信度，支持联合利用这些互补标签的多任务学习方法。

描述性分析揭示了若干对下游建模具有启示意义的模式。中性情绪在语料库中占主导地位，这反映了广播语音的专业语域；与此同时，方言内容是最大的类别。单人独白片段占比较高，这表明广播语音可能包含独特的地域线索，有助于方言识别。然而，各城市间显著的地理分布不均衡，意味着需要采用采样策略或加权损失函数，以确保模型性能公平。音频质量分析进一步揭示了信噪比存在系统性的地理差异，这关系到如何区分声学劣化与方言特征。通过发布这一语料库，并提供有关数据收集和标注规范的透明文档，我们为研究界推进基于音频的细粒度方言归属判定奠定了基础。潜在应用不仅限于方言识别，还包括社会语言学研究、领域偏移下的鲁棒性评估，以及向电话通信和在线视频等相关媒介进行迁移学习。可重用的流程和标注指南也

有助于未来扩展和社区贡献。详细的实验基准评测留待后续研究开展。

参考文献

[1] Serry Sibaee、Omer Nacar、Yasser Al-Habashi、Adel Ammar 和 Wadii Boulila。SHAMI-MT：叙利亚阿拉伯方言与现代标准阿拉伯语双向机器翻译系统，2025年。[2] Jihad Zahir。IADD：综合阿拉伯方言识别数据集。*Data in Brief*，40：107777，2021年。[3] Omer Nacar、Serry Taiseer Sibaee、Samar Ahmed、Safa Ben Atitallah、Adel Ammar、Yasser Alhabashi、Abdulrah- man S Al-Batati、Arwa Alsehibani、Nour Qandos、Omar Elshehy 等。迈向包容性阿拉伯语大语言模型：阿拉伯语大语言模型评测中的文化对齐基准。载于第一届低资源语言大语言模型研讨会论文集，第387–401页，2025年。[4] Muhammad Abdul-Mageed、Amr Keleg、AbdelRahim Elmadany、Chiyu Zhang、Injy Hamed、Walid Magdy、Houda Bouamor 和 Nizar Habash。NADI 2024：第五届细粒度阿拉伯方言识别共享任务。载于阿拉伯语自然语言处理会议论文集，2024年。arXiv预印本 arXiv:2407.04910。[5] Bashar Talafha、Hawau Olamide Toyin、Peter Sullivan、AbdelRahim A Elmadany、Abdurrahman Juma、Amirbek Djanibekov、Chiyu Zhang、Hamad Alshehhi、Hanan Aldarmaki、Mustafa Jarrar 等。NADI 2025：首届多方言阿拉伯语语音处理共享任务。载于第三届阿拉伯语自然语言处理会议：共享任务论文集，第720–733页，2025年。[6] Su won Shon、Ahmed Ali、Younes Samih、Hamdy Mubarak 和 James Glass。ADI17：细粒度阿拉伯方言识别数据集。载于*IEEE*声学、语音与信号处理国际会议（*ICASSP*）论文集，第8244–8248页，2020年。[7] Karima Meftouh、Salima Harrat、Salma Jamoussi、Mourad Abbas 和 Kamel Smaili。PADIC上的机器翻译实验：阿拉伯方言平行语料库。载于 Hai Zhao（编），第*29*届亚太语言、信息与计算会议论文集，第26–34页，中国上海，2015年10月。[8] Amal Nayouf、Tymaa Hammouda、Mustafa Jarrar、Fadi Zaraket 和 Mohamad-Bassam Kurdy。Nâbra：带有形态标注的叙利亚阿拉伯方言。载于 Hassan Sawaf、Samhaa El-Beltagy、Wajdi Zaghouani、Walid Magdy、Ahmed Abdelali、Nadi Tomeh、Ibrahim Abu Farha、Nizar Habash、Salam Khalifa、Amr Keleg、Hatem Haddad、Imed Zitouni、Khalil Mrini 和 Rawan Almatham（编），*ArabicNLP 2023*论文集，第12–23页，新加坡（线上线下混合），2023年12月。计算语言学协会。[9] Salima Medhaffar、Fethi Bougares、Yannick Estève 和 Lamia Hadrich-Belguith。突尼斯方言情感分析：语言资源与实验。载于 Nizar Habash、Mona Diab、Kareem Darwish、Wassim El-Hajj、Hend Al-Khalifa、Houda Bouamor、Nadi Tomeh、Mahmoud El-Haj 和 Wajdi Zaghouani（编），第三届阿拉伯语自然语言处理研讨会论文集，第55–61页，西班牙瓦伦西亚，2017年4月。计算语言学协会。[10] Omer Nacar、Anis Koubaa、Serry Sibaee、Yasser Al-Habashi、Adel Ammar 和 Wadii Boulila。GATE：用于提升语义文本相似度的通用阿拉伯语文本嵌入，采用套娃式表征学习与混合损失训练。*arXiv*预印本 *arXiv:2505.24581*，2025年。[11] Alexandra Canavan、George Zipperlen 和 David Graff。CALLHOME埃及阿拉伯语语音。语言数据联盟，1997年。LDC97S45。[12] Peter Bell、Mark JF Gales、Thomas Hain、Jonathan Kilgour、Pierre Lanchantin、Xunying Liu、Andrew McParland、Steve Renals、Oscar Saz、Mirjam Wester 等。MGB挑战赛：多体裁广播媒体识别评测。载于*2015*年*IEEE*自动语音识别与理解研讨会（*ASRU*），第687–693页。IEEE，2015年。[13] Ahmed Ali、Peter Bell、James Glass、Yacine Messaoui、Hamdy Mubarak、Steve Renals 和 Yifan Zhang。MGB-2挑战赛：阿拉伯语多方言广播媒体识别。载于*2016*年*IEEE*口语语言技术研讨会（*SLT*），第279–284页，2016年。[14] Ahmed Ali、Stephan Vogel 和 Steve Renals。真实环境下的语音识别挑战：阿拉伯语MGB-3。载于*2017*年*IEEE*自动语音识别与理解研讨会（*ASRU*），第316–322页，2017年。[15] Ahmed Ali、Suwon Shon、Younes Samih、Hamdy Mubarak、Ahmed Abdelali、James Glass、Steve Renals 和 Khalid Choukri。MGB-5挑战赛：阿拉伯方言语音识别与方言识别。载于*2019*年*IEEE*自动语音识别与理解研讨会（*ASRU*），第1026–1033页。IEEE，2019年。

[16] Hamdy Mubarak、Amir Hussein、Shammur Absar Chowdhury 和 Ahmed Ali。QASR：QCRI半岛电视台语音资源——大型阿拉伯语语音标注语料库。载于 Chengqing Zong、Fei Xia、Wenjie Li 和 Roberto Navigli（编），计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第2274–2285页，线上，2021年8月。计算语言学协会。[17] Fadi Biadsy、Julia Hirschberg 和 Nizar Habash。采用音系搭配建模进行口语阿拉伯语方言识别。载于*EACL 2009*闪米特语言计算方法研讨会论文集，Semitic '09，第53–61页，美国，2009年。计算语言学协会。[18] Nour Qandos、Serry Sibaee、Samar Ahmad、Omer Nacar、Adel Ammar、Wadii Boulila 和 Yasser Alhabashi。Anplers参与iqraeval共享任务：将Whisper-large-v3调整为语音转音素模型，用于检测《古兰经》诵读中的发音错误。载于第三届阿拉伯语自然语言处理会议：共享任务论文集，第464–468页，2025年。[19] Kees Versteegh。阿拉伯语。爱丁堡大学出版社，2014年。[20] Raya Kalaldeh。约旦语 speakers 所说现代标准阿拉伯语元音的声学分析。《阿拉伯语—英语研究国际期刊》，18(1)，2018年。[21] Emad Abdul-Latif。Bassiouney，Reem（2009），《阿拉伯语社会语言学》。爱丁堡大学出版社：爱丁堡；第 i–xiv 页。《阿拉伯语言学国际期刊》，2(2)：154–158，2016年。[22] Alvan R Feinstein 和 Domenic V Cicchetti。高度一致性但卡帕系数偏低：一、两个悖论带来的问题。《临床流行病学杂志》，43(6)：543–549，1990年。