

《AI终端语音测试数据集建设指南》 标准编制说明

标准起草小组

1. 标准范围

本文件规定了AI终端语音测试数据集建设过程中的总体原则、多场景覆盖设计、口音与说话人多样性设计、标注规范、质量控制体系及安全隐私要求。

本文件适用于AI终端语音测试数据集的规划、采集、标注、质检与交付。

本文件不适用于专门用于模型训练的数据集建设。训练数据集的安全底线、通用元数据标签、轻量化格式原则等，应遵循《人工智能终端模型训练数据集建设通用指南》。

2. 工作简况

3. 标准编制原则和确定标准主要内容

本标准依据《标准化工作导则 第1部分：标准化文件的结构和起草规则》（GB/T 1.1-2020）编制。本标准的主要内容基于AI终端语音交互产业测试数据评价体系不统一的痛点和技术现状确定，结合起草单位技术实践经验，确保内容的科学性和行业适配性。

本标准核心内容包括：

总体要求：明确代表性、多样性、均衡性等建设原则，规定音频、文本、元数据的数据构成要素，并界定基础功能、性能基准、大规模压力的规模要求及单句时长分布指标，确保数据集满足不同深度的测试需求。

场景多样性设计：明确安静、常规噪声、复杂干扰三类声学环境要求，规定近、中、远场空间距离及角度覆盖，涵盖终端播放态、通话态等回声消除干扰场景，以及唤醒、打断、重叠语音等全模式交互场景。

口音与说话人多样性设计：明确标准音、方言口音、外籍口音的覆盖层级与比例，规定儿童至老年全年龄段、不同性别及特殊生理特征的说话人样本分布，确保测试集对各类用户群体的泛化能力与公平性。

标注规范：明确文本转写、事件边界标注的准确性要求，规定填充词保留、错读标注等细节处理规则，定义结构化元数据字段格式。

质量控制：明确音频质量、标注质量、数据完整性三类指标体系，规定采集、标注、验收全流程质控机制，引入双盲标注与基线系统“试跑”验收机制，确保数据集具备足够的区分度与鲁棒性。

安全与交付 明确个人信息脱敏、生物特征注销等隐私保护要求，规定数据加密存储与传输规范，定义标准化的文件命名规则与交付格式，确保数据集的合规流通与高效利用。

4. 主要试验(或验证)的分析、综述报告

无。

5. 标准在起草过程中遇到的问题及解决办法；重大分歧意见的处理经过和依据；有无重要技术问题需要说明

在本文件的修订过程中，无重大分歧意见和技术问题。

6. 与国外标准的关系：包括：采用国际标准和国外先进标准的程度，与国外标准主要技术内容的差异

该项目没有对应的国际标准或国外先进标准。

7. 修订标准时，说明与标准前一版本的重大技术变化，并列岀所涉及的新、旧版本的有关章条(可引用标准前言的内容)；废止/代替现行有关标准的建议

不涉及。

8. 说明标准与其他标准或文件的关系(可引用标准前言的内容)，特别是与有关的现行法律、法规和强制性国家标准的关系

符合现行法律、法规要求。

9. 标准作为强制性标准或推荐性标准的建议

建议本文件作为推荐性标准。

10. 贯彻国家标准的要求和措施建议(包括组织措施、技术措施、过渡办法等内容)；标准发布后，对国内外业界可能产生的影响

建议本文件作为推荐性标准发布实施。

本标准旨在规范AI终端语音测试数据集建设，规定了AI终端语音测试数据集的场景覆盖要求、口音多样性规范、数据标注规则及质量控制流程，明确数据集建设的基线与评价指标。

随着智能音箱、车载语音助手等AI终端设备的爆发式增长，语音交互已成为人机交互的核心入口。然而，当前行业缺乏统一的测试数据集建设标准，导致测试数据规模虚高、分布失真、场景覆盖不全等问题频发，造成不同厂商间的评测结果不可比、系统鲁棒性验证不充分。制定本标准旨在解决测试数据质量偏低与评价体系碎片化难题，为AI终端性能评估提供“统一标尺”。

通过规范场景覆盖要求、口音多样性规范、数据标注规则及质量控制流程，指导建设者构建AI终端语音测试数据集。实施本标准有助于统一行业AI终端语音测试数据集标准，提升AI终端在复杂环境下的交互成功率和用户体验。

11. 标准是否涉及知识产权的情况说明；如标准中含有自主知识产权，说明产品研发程度、产业化基础及进程

本文件未涉及。

12. 其他应予说明的事项

本文件未涉及。