

团 体 标 准

T/ISC XXX—XXXX

AI 终端语音测试数据集建设指南

Construction guidelines for speech test dataset of AI terminals

在提交反馈意见时，请将您知道的相关专利与支持性文件一并附上。

（征求意见稿）

XXXX—XX—XX 发布

XXXX—XX—XX 实施

中 国 互 联 网 协 会 发 布

目 次

前 言	IV
引 言	VI
AI 终端语音测试数据集建设指南	1
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
3.1 AI 终端 AI terminal	1
3.2 语音测试数据集 speech test dataset	1
3.3 多场景覆盖 multi-scenario coverage	1
3.4 口音多样性 accent diversity	1
4 符合和缩略语	1
5 总体要求	2
5.1 建设原则	2
5.2 数据集构成	2
5.3 规模要求	2
6 多场景覆盖设计	3
6.1 声学环境分类	3
6.2 空间与距离设计	3
6.3 终端干扰态设计	3
6.4 语音交互模式覆盖	3
7 口音与说话人多样性设计	3
7.1 方言与口音覆盖	3
7.2 生理与人口统计学特征	4
8 标注规范	4
8.1 文本转写规范	4
8.2 事件与边界标注	4
8.3 元数据规范	4
9 质量控制	5
9.1 质量指标体系	5
9.2 质量控制流程	5
10 安全与隐私保护	6
11 数据集交付格式	6
11.1 文件命名规则	6
附 录 A （资料性）	7
A.1.1 AI 终端语音测试数据集难度分级参考表	7
参 考 文 献	8

前 言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国互联网协会提出并归口。

本文件起草单位：

本文件主要起草人：

引言

随着人工智能技术的快速发展，AI终端设备（如智能手机、智能音箱、车载系统、可穿戴设备等）的语音交互功能已成为用户体验的核心组成部分。语音测试数据集作为评估AI终端语音识别、语义理解、语音合成等能力的基础资源，其建设质量直接影响测试结果的准确性和可靠性。

当前，AI终端语音测试面临场景覆盖不全、口音代表性不足、标注标准不统一、质量控制体系不完善等问题。为指导行业规范建设高质量的语音测试数据集，特制定本文件。本文件从多场景覆盖、口音多样性、标注规范及质量控制四个维度，提供可操作的构建指南，支撑AI终端语音能力的全面、客观、公正测试。

本文件与《人工智能终端模型训练数据集建设通用指南》协同制定。训练阶段数据集建设遵循后者；本文件仅规范语音能力客观评测用测试集，两者在代表性、多样性、均衡性、边界对抗性、可追溯性等建设原则及安全隐私底线上保持一致。

AI 终端语音测试数据集建设指南

1 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 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。

GB/T 36464（所有部分） 信息技术 智能语音交互系统

GB/T 21023-2007 中文语音识别系统评测规范

GB/T 35273—2020 信息安全技术 个人信息安全规范

GB/T 32625—2016 信息技术 中文语音合成系统通用技术规范

GB/T 36625.3—2021 智慧城市 数据融合 第3部分：数据采集规范

3 术语和定义

下列术语和定义适用于本文件。

3.1 AI 终端 AI terminal

具备麦克风阵列、语音处理芯片及人工智能算法，能够通过语音交互实现特定功能的终端设备。

3.2 语音测试数据集 speech test dataset

为评估AI终端语音交互系统性能（如唤醒率、识别准确率、响应时延等），而专门构建的具有明确边界条件和客观评价标准的标准化语音数据集合。

3.3 多场景覆盖 multi-scenario coverage

语音测试数据集涵盖AI终端实际使用中可能遇到的各种声学环境、使用场景和交互模式的程度。

3.4 口音多样性 accent diversity

语音测试数据集中包含的不同地域口音、社会口音及非母语口音的丰富程度。

4 符合和缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

AEC：声学回声消除（Acoustic Echo Cancellation）

dB: 分贝 (Decibel)

dBFS: 满量程相对分贝 (Decibels Full Scale)

JSON: JS对象简谱 (JavaScript Object Notation)

MAE: 平均绝对误差 (Mean Absolute Error)

PII: 个人身份信息 (Personally Identifiable Information)

RT60: 混响时间 (Reverberation Time 60dB)

SNR: 信噪比 (Signal-to-Noise Ratio)

5 总体要求

5.1 建设原则

本文件建设原则与《人工智能终端模型训练数据集建设通用指南》第5章总体框架中的代表性、多样性、均衡性、边界对抗性、可追溯性原则保持一致，并针对语音评测场景进一步细化。

AI终端语音测试数据集的建设应遵循以下原则：

5.1.1 代表性原则

样本分布应真实反映目标用户群体及实际使用环境的统计学特征。

5.1.2 多样性原则

样本分布应覆盖不同用户群体、使用环境、交互方式。

5.1.3 均衡性原则

各类场景、口音、信噪比、性别的样本比例应避免严重倾斜，防止测试偏差。

5.1.4 边界对抗性原则

应包含易混淆词汇、极端噪声、远场极限距离等高难度样本，以检验系统的鲁棒性。

5.1.5 可追溯性原则

每条语音数据均应配备完整的元数据，确保测试结果异常时可精准定位数据缺陷。

5.2 数据集构成

AI终端语音测试数据集应包含以下核心组成部分：

- a) 音频数据：原始语音录音文件，采样率需 ≥ 16 kHz，量化位数需 ≥ 16 bit；
- b) 文本标注：语音内容的准确转写文本，包含标点符号和特殊标记；
- c) 元数据：包括说话人信息（性别、年龄段、口音类型）、录音环境信息（场景类型、信噪比）、设备信息（终端类型、麦克风规格）等；
- d) 切分信息：语音段落的起止时间戳、有效语音边界等。

5.3 规模要求

AI终端语音测试数据集的规模应根据测试目的确定：

- a) 基础功能测试：有效语音时长不少于100小时，句数不少于40000句；
- b) 性能基准测试：有效语音时长不少于500小时，句数不少于200000句；
- c) 大规模压力测试：有效语音时长不少于2000小时，句数不少于800000句。

- d) 单条语音时长要求：单条语音有效时长应控制在0.5秒至60秒之间。为模拟真实交互场景，数据集应具备合理的时长分布，其中短语音（0.5秒~4秒）占比宜保持在40%~60%，中长语音（4秒~20秒）占比宜保持在30%~50%，超长语音（20秒以上）占比不宜超过10%。

6 多场景覆盖设计

6.1 声学环境分类

声学环境应划分为以下三类，并明确比例分配（推荐比例 4：4：2）：

- 安静室内环境：混响时间（RT60）小于0.4s，信噪比（SNR） $\geq 20\text{dB}$ 。包含：卧室、办公室、书房等。
- 常规噪声环境：存在稳态或准稳态噪声，信噪比（SNR）在0dB~15dB之间。包含：客厅（电视声）、驾驶舱（风噪+发动机噪声）、咖啡厅（非重叠人声）。
- 复杂干扰环境：存在非稳态噪声或强混响，信噪比（SNR）在-5dB~5dB之间。包含：马路交叉口（喇叭声+施工声）、KTV包厢、厨房（抽油烟机+切菜声）

6.2 空间与距离设计

针对搭载麦克风阵列的AI终端，需设置梯次距离测试集：

- 近场：0.2m ~ 0.5m（手持/佩戴状态），样本占比20%。
- 中场：1.0m ~ 2.0m（桌面/床头状态），样本占比50%。
- 远场：3.0m ~ 5.0m（跨房间/远距离状态），样本占比30%。
- 角度覆盖：在3m及以上距离，需覆盖水平角0°、45°、90°、135°、180°及俯仰角 $\pm 15^\circ$ 、 $\pm 45^\circ$ 。

6.3 终端干扰态设计

AI终端语音测试数据集需包含AI终端自身发声带来的回声消除（AEC）测试场景：

- 播放态：终端以中等音量（约65dB）播放音乐或播客时录入语音。
- 通话态：终端处于免提通话状态时录入语音。
- 静默态：终端待机状态的常规录入。

6.4 语音交互模式覆盖

AI终端语音测试数据集应覆盖以下语音交互模式：

- 唤醒词交互：包含唤醒词及唤醒词+指令的连续语音。
- 近场单轮交互：近距离单轮问答式语音。
- 近场多轮交互：连续多轮对话语音。
- 远场交互：距离设备1 m以上的语音交互。
- 打断交互：在设备播报过程中插入的语音指令。
- 重叠语音：多人同时说话场景。

7 口音与说话人多样性设计

7.1 方言与口音覆盖

普通话测试集应以《普通话水平测试大纲》为基础，按以下层级设计口音分布：

- a) 标准音：一级乙等及以上水平，无明显地方口音，占比不低于30%。
- b) 优势方言口音：包含但不限于粤语口音、四川话口音、东北话口音、河南话口音、吴语口音（温州/上海）、闽南语口音，各地方言区口音合计占比55%~65%，单一口音占比5%~10%。
- c) 外籍/少数民族口音：包含带有典型外籍特征的汉语口音（如日式、韩式、俄式汉语发音），占比5%~10%。

7.2 生理与人口统计学特征

为保证算法的公平性，说话人分布应满足：

- a) 年龄段：儿童（6-12岁）10%、青年（13-35岁）40%、中年（36-55岁）30%、老年（56岁以上）20%。*注：重点覆盖老年人发声器官退化导致的漏风音、颤音特征。*
- b) 性别：男女性别比例控制在 1:1 左右。
- c) 特殊生理特征：包含轻微感冒鼻音、嘶哑声、口吃等异常发声样本，占比3%~5%。

8 标注规范

8.1 文本转写规范

转写应忠实于语音内容，不得添加、遗漏或修改语义。

- a) 填充词：保留“啊”、“呃”等非语义填充词，建议使用特殊标识（如<fill>或[uh]）进行标注，以还原真实语音流。
- b) 异读与错读保留：说话人发音错误（如将“公司”读作“公私”）或方言词汇，应按实际发音使用同音字/近音字转写，并在后方加注〔错读〕标签。
- c) 侧信道标注：对被环境噪声完全掩蔽、人耳无法辨识的语音片段，使用〔不可辨〕标签，禁止主观猜测。

8.2 事件与边界标注

针对语音唤醒和端点检测（VAD）测试，需进行毫秒级时间戳标注：

- a) 唤醒词边界：精确标注唤醒词的起始时间（精确到音素起始点）和结束时间。
- b) 噪声事件边界：对突发噪声（如关门声、狗叫声）标注起止时间及类型。
- c) 重叠语音边界：标注第三者介入说话的起止时间。

8.3 元数据规范

标注文件应采用结构化格式存储，推荐使用JSON或XML格式，必选字段见表1。

表 1 语音测试数据集元数据必选字段

字段名	数据类型	描述说明
file_id	String	音频文件唯一标识符
text	String	规范化转写文本
speaker_id	String	说话人脱敏ID
gender	Enum	选项: male, female, other
age_group	Enum	选项: child, youth, middle, senior
accent	String	口音/方言类型描述
scene_type	Enum	选项: quiet, normal_noise, complex_noise
snr_db	Float	估计信噪比, 保留1位小数
distance_m	Float	嘴部距麦克风距离 (米)
device_model	String	采集设备型号 (如使用终端自身麦克风则标明终端型号)
sample_rate	Int	采样率 (如16000)
wake_word_boundaries	Array	唤醒词起止时间数组, 如 [1.02, 1.45]

9 质量控制

9.1 质量指标体系

语音测试数据集质量应通过以下指标进行评估：

9.1.1 音频质量指标

- a) 采样率一致性：100%符合规定采样率；
- b) 音量一致性：平均音量在-20 dBFS~-10 dBFS范围内；
- c) 信噪比达标率：≥90%的样本信噪比符合场景要求；
- d) 截幅率：音频截幅样本占比≤1%。

9.1.2 标注质量指标

- a) 字准确率：标注文本与参考标准的一致性≥99%；
- b) 时间对齐精度：边界误差≤50 ms的占比≥95%；
- c) 标注一致性：多人标注一致性≥98%；
- d) 漏标率：漏标内容占比≤0.5%。

9.1.3 数据完整性指标

- a) 元数据完整率：100%；
- b) 文件完整性：无损坏、无丢失；
- c) 格式合规率：100%。

9.2 质量控制流程

9.2.1 采集前质量控制

- a) 设备校准：使用声级计对采集终端及参考麦克风进行声压级校准，确保信噪比标注的基准一致。

- b) 脚本审查：由语言学专家对测试文本集进行“最小对立体”和“音素覆盖率”审查，确保测试集无语义盲区。

9.2.2 采集后音频质检

- a) 自动化盲筛：通过VAD算法剔除有效语音时长小于1秒或断句超过5秒的废音。
- b) 频域检验：剔除因设备削波导致的频谱截顶失真音频，或因直流偏移导致的低频异常音频。
- c) 一致性抽检：随机抽取5%的音频，人工复核其实际声学环境是否与元数据标注一致。

9.2.3 标注质量控制

- a) 双盲标注：所有文本与边界标注必须由两名独立标注员完成。
- b) 一致性计算：使用Cohen's Kappa系数评估文本一致性（要求 $\kappa > 0.85$ ），使用平均绝对误差（MAE）评估边界时间戳一致性（要求 $MAE < 50ms$ ）。
- c) 仲裁机制：对不一致样本引入第三名高级标注员进行仲裁，仲裁结果作为最终标签。

9.2.4 验收测试机制

数据集交付前，应使用至少3款主流商用AI终端作为基线系统进行“试跑”测试。若基线系统在某特定子集上的准确率极值差小于5%（即所有系统表现趋于一致），或平均准确率趋近于0%或100%，说明该子集缺乏难度梯度或区分度，需重新调整数据集分布。

10 安全与隐私保护

- a) 数据脱敏：音频文本中若包含身份证号、电话号码、家庭住址等个人敏感信息（PII），必须在转写文本中进行掩码处理（如：138***5678），或直接重录。
- b) 生物特征注销：严禁在元数据中保存说话人真实姓名、联系方式。说话人声纹应被视为生物特征信息，提供明确的注销机制。
- c) 授权合规：数据采集前必须取得数据主体的明示同意，授权范围应明确包含“用于人工智能算法测试与评估”。
- d) 存储加密：传输与存储过程应符合GB/T 35273—2020要求，通过算法对原始音频进行加密落盘。
- e) 测试集严禁中心化汇聚未脱敏的原始敏感音频，声纹信息按生物特征信息管理。

11 数据集交付格式

11.1 文件命名规则

音频与标注文件应采用强关联命名，例如：

[场景码]_[口音码]_[信噪比]_[说话人ID]_[序号].wav

示例：Complex_Kitchen_SC_10dB_SP001_0001.wav（代表：复杂干扰环境-厨房、四川口音、10分贝信噪比、1号说话人、第1条）。

见附录A

附 录 A
(资料性)

A. 1. 1 AI终端语音测试数据集难度分级参考表

表 B. 1 AI 终端语音测试数据集难度分级参考表

难度等级	信噪比范围	距离	交互状态	适用测试阶段
L1（简单）	> 15dB	< 0.5m	静默态	回归测试、基础功能验收
L2（中等）	5dB ~ 15dB	0.5m ~ 2.0m	播放态(低音量)	日常用户体验评估
L3（困难）	0dB ~ 5dB	2.0m ~ 3.5m	播放态(高音量)	鲁棒性专项测试
L4（极限）	< 0dB	> 3.5m	强干扰态(如重叠语音)	算法能力边界探索

参 考 文 献

- [1] GB/T 1.1-2020 标准化工作导则 第1部分：标准化文件的结构和起草规则[S]. 北京：中国标准出版社，2020.
- [2] GB/T 7713.2-2022 学术论文编写规则[S]. 北京：中国标准出版社，2022.
- [3] GB/T 5271.29-2006 信息技术 词汇 第29部分：人工智能 语音识别与合成[S]. 北京：中国标准出版社，2006.
- [4] GB/T 21023-2007 中文语音识别系统评测规范[S]. 北京：中国标准出版社，2007.
- [5] GB/T 36464（所有部分） 信息技术 智能语音交互系统[S]. 北京：中国标准出版社.
- [6] 中国语音产业联盟. 中文语音合成系统评测规范[S]. 2018.
- [7] 中国信息通信研究院. 智能音箱语音识别评测方法[S]. 2020.
- [8] 中国人工智能产业发展联盟. 智能语音识别服务系统评估规范[S]. 2021.
-