

训练数据需求说明

1、录音文本

- (1) 按照项目需求提供所有录音数据对应的命令行文本；
- (2) 命令行音节长度建议在 4-8 个音节，中文基本一字为一音节，如打开空调为 4 个音节，其他语种以一个元音单位为一个音节，如英文 'air -con di tion er' 为 5 个音节；
- (3) 小语种命令行需提供文本对应的中文释义或英文释义；
- (4) 小语种中包含的外来语应尽量提供其对应语种的拼写形式，
如：中文中“USB”可写为“you1 ai5 si5 bi4”，
韩语中“time”可写为“타이머”，
越南语“massage”可写为“Mát xa”；
- (5) 将文本中所有的阿拉伯数据转写成录音语种对应的文字格式，
如：中文“1 小时”需改写为“一小时”，
英语“1 hour”需改写为“one hour”，
韩语“1 시간”需改写为“한 시간”；

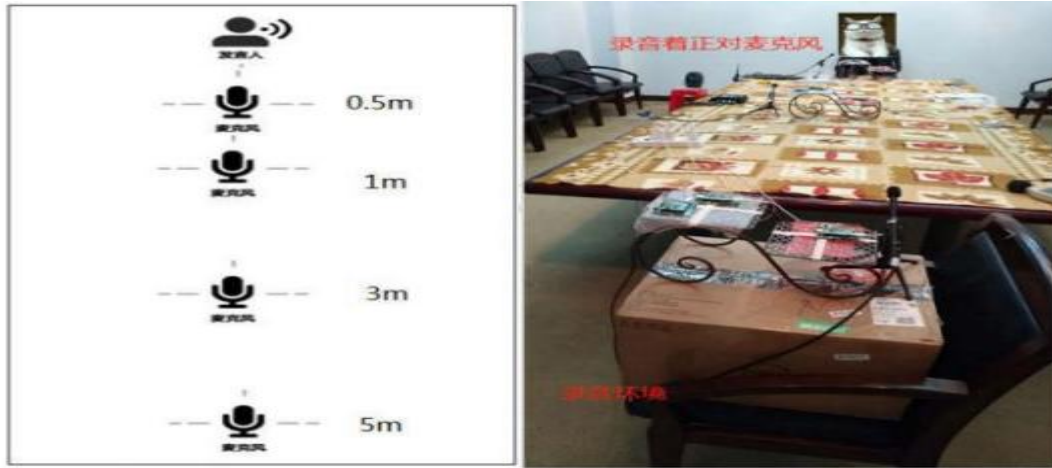
2、录音人员及流程

- (1) 训练集音频录音人建议主要选取年龄在 18-60 的人群（以贵司产品的用户群体做参考），
- (2) 小语种开发训练的数据采集，建议采集人次不低于 50 人次；若是喉音、颤音等特殊发音较多语种，则建议录制人数不低于 100 人次；
- (3) 中文项目增强训练的数据采集，建议采集人次不低于 150 人次；若是新开发领域的项目，则建议录制人数不低于 300 人次；
- (4) 以当地标准口音录制，尽量不要带太重的地方口音；
- (5) 录音人文字识别朗读无困难，且发音流畅，避免卡顿或一个字一个字发音的情况；
- (6) 相同命令行至少录制 3 遍，正常语速 1 遍，快语速 2 遍；
- (7) 唤醒词每人次录制不少于 10 遍；
- (8) 单个命令行录音过程不中断；
- (9) 语速平稳，并避免音量参差不齐；
- (10) 建议性别比例 1:1；
- (11) 发音人面对最近 mic 分贝在 65-80dB 左右；
- (12) 采集时记录发音人的性别、年龄和籍贯信息，记录录音使用的录音设备及型号(如罗兰 R44 录音机、或华为 mate50 等)，并与其音频存储的文件一一对应；
- (13) 提供与音频发音顺序对应的录音文本；

3、采集设备及环境要求(高保真录音机)

- (1) 最好适宜家居环境，现场混响范围在 0.3-0.6；
- (2) 房间底噪在 35-45dB；

- (3) 最好使用罗兰高保真录音机，根据现场环境调试设备增益和灵敏度；
- (4) mic 选取舒伯乐 ECM999 麦克风四个，对应不同距离 mic 采集音频；
- (5) 麦克风放在距离发声源 0.5m、1m、3m、5m 位置，如有必要可同时放置不同角度对应的 mic 进行录音；
- (6) 现场环境保持安静，无明显的干扰噪音，现场环境搭建如下：



4、采集设备及环境要求(手机/电脑/高保真录音笔)

- (1) 手机/电脑录音需保证较高的音频质量，采样率不低于 44.1k，且避免传输时造成音质受损；
- (2) 使用手持高保真录音笔，选择 44.1k 采样率单声道进行录制；
- (3) 使用手机/电脑/手持高保真录音笔时，将收音设备放在距离发音人 3m 的位置；
- (4) 最好适宜家居环境，录音环境混响尽量小，建议 0.3-0.6；
- (5) 房间底噪在 35-40dB；
- (6) 吐字清晰，保证声音不要削顶，减少数据损失；
- (7) 每个命令词单独录制，则每句话前后间隔 1-2s；
- (8) 现场环境保持安静，无明显的干扰噪音；
- (9) 音频存储格式 wav、采样率不低于 44.1K；

5、数据存储注意事项

- (1) 命令词较少时，建议每条命令词存储为 1 个文件，并提供对应的文本；
- (2) 命令词较多时，建议使用启英泰伦录音工具搭配高保真录音机进行录制，可实现后期数据的批量切分；
- (3) 启英泰伦录音工具备建方法请联系启英泰伦数据组工作人员。
- (4) 严格按照文档示例命名、存储。