

远场语音助手语音数据库说明书

AISHELL-ASR0070



希尔贝壳
A I S H E L L

Copyright

目录

1 产品概述.....	3
2 产品目录结构.....	3
2.1 目录结构.....	3
2.2 命名规则.....	4
3 文本设计.....	4
4 录制环境.....	4
4.1 录制环境.....	4
4.2 录制设备.....	5
4.3 录制方法.....	5
5 标注转写规范.....	6
6 发言人信息.....	7
7 版权声明.....	7

1 产品概述

AISHELL-ASR0070 远场语音助手语音数据库共 6200 小时，邀请 450 名来自中国南北方区域发言人参与，以北方为主。录音语言，中文；录音地区，中国。录音文本内容为语音助手交互内容。录制过程在真实家居环境中，客厅、卧室、厨房分别设置 5、4、3 个录音位，录制使用圆形 16 路 PDM 麦克风阵列录音板(16KHz、16bit；6083H)做远讲拾音和高保真麦克风做近讲拾音(44.1KHz、16bit；117H)。

此数据库经过专业语音校对人员转写标注，并通过严格质量检验，文本正确率在 98% 以上。可用于声纹识别、语音识别等研究使用。

2 产品目录结构

2.1 目录结构

数据目录结构	
数据目录结构	
AISHELL-ASR0070.pdf	(数据库简介)
└─SPEECHDATA	(音频文件夹)
─P0001	(发言人文件夹)
P0001B01	(录音位文件)
P0001B01C01	(录音位信号通道文件)
P0001B01C01W0001.wav	(音频文件)
└─DOC	(文本文件夹)
spk_info.xlsx	(发言人信息)
content.txt	(标注内容)
wav_list.txt	(音频列表)
ASR0070_Noise_info.xlsx	(场景噪音信息列表)

图表 2-1-1 目录结构

2.2 命名规则

`<speakerID>\<deviceID>\<deviceID_channelID>\<wavID>.wav`

e.g. P0001/P0001B01/P0001B01C01/ P0001B01C01W0001.wav

名称	内容	备注
P0001	0001-0500	发言人编号
P0001B01	<meetingID>_B01 & 02 & 03 & 04 & 05	设备编号
P0001B01C01	<meetingID>_<deviceID>_C01~16	信道编号
P0001B01C01W0001.wav	<meetingID>_<deviceID>_<deviceID_channelID>_W0001~	音频编号

图表 2-2-1 路径命名规则

3 文本设计

本数据库文本定义为语音助手交互内容。由海量真实语音助手文本生成，内容包含生活、娱乐、工作等 38 个领域，并经过人工筛选校对。每人按照 80 句随机对应一个场景，共 240 句。例如录音人 P0001：（参照场景噪音信息表

序号	场景类型	噪音类型
0001~0080	客厅	安静
0081~0160	厨房	安静
0161~0240	厨房	油烟机

4 录制环境

4.1 录制环境

1. 录制场景为真实家居场景，覆盖客厅、卧室、厨房。
2. 录音人站在指定位置。

4.2 录制设备

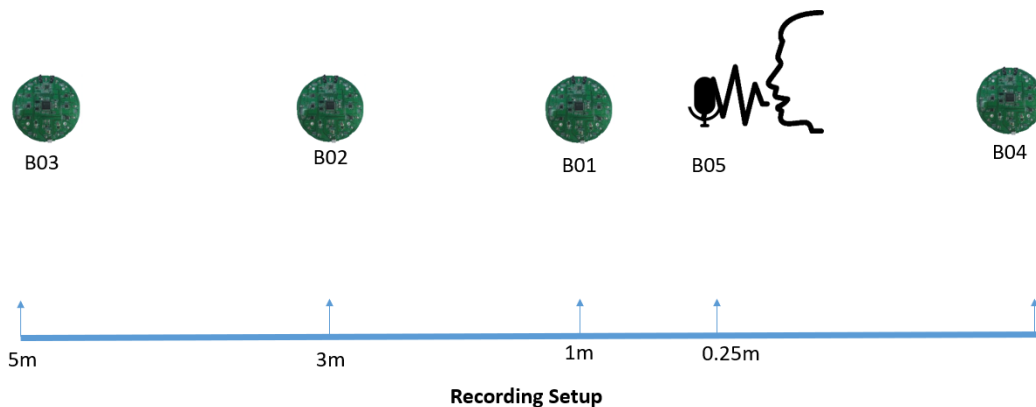
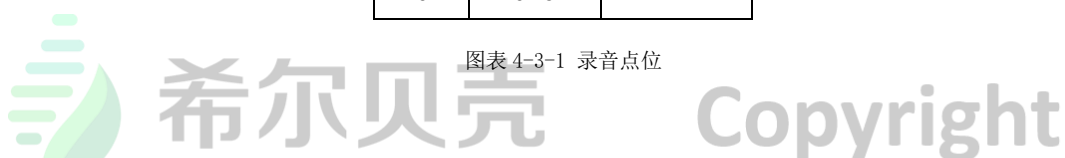
录制设备采用 16 路 PDM 麦克风圆型录音开发板。

4.3 录制方法

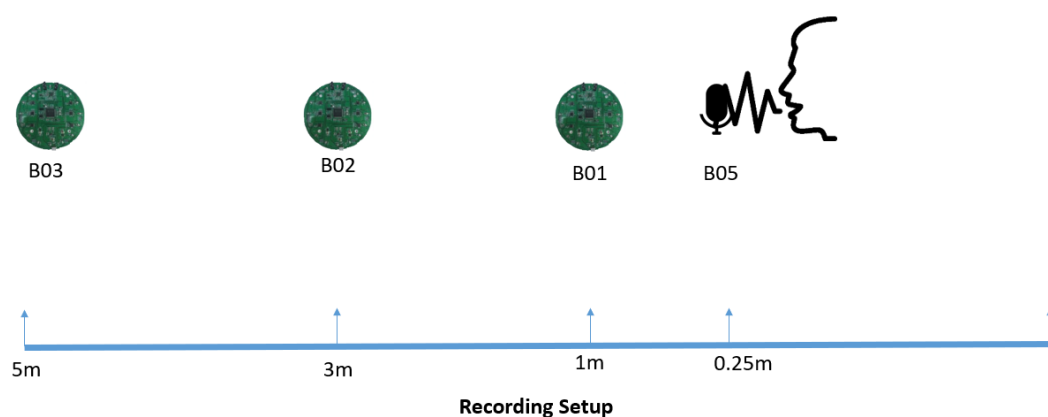
在固定客厅、卧室、厨房场景环境中分别在 2、3、4 个不同位置（相对录音人）放置拾音设备，录音点位的信息如下表：

编号	距离 (m)	高度 (m)
B01	1	1
B02	3	1
B03	5	1
B04	-1.5	1
B04	0.25	~

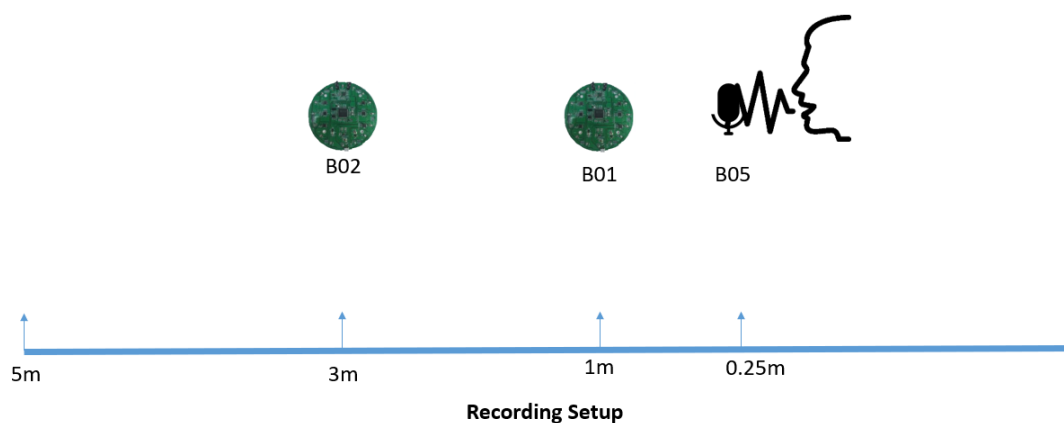
图表 4-3-1 录音点位



图表 4-3-2 客厅录制示意图



图表 4-3-3 卧室录制示意图



图表 4-3-3 厨房录制示意图

5 标注转写规范

数据转写人员根据所听到的音频写出内容，力求使文本内容与音频发音内容保持一致。

准则如下：

- 1) 转写的内容必须和听到的语音完全一致，不能多字、少字、错字。
- 2) 数字要转写为汉字形式，如“一二三”，而不是“123”。注意区分“一”和“幺”，“二”和“两”。
- 3) 句中出现的英文按照发音写出单词，如“thank you”。按拼读朗读的字母，需转写成大写字母加空格的形式。如，“NBA”、“UFO”。

- 4) 句中包含的符号，按实际发言人发音转写。如“三 W 点 百度 点 com”。没有发音的符号，需要删掉。品牌名称，专有名称等按照实际惯用格式转写，如“QQ 空间”、“iPhone”、“喜马拉雅”。
- 5) 标注内容的完整性要与实际发音一致，不得删减。

6 发言人信息

发言人信息记录内容包括录音编号、性别、年龄、籍贯区域。如图表所示：

录音编号	性别	年龄	区域
P0001	女	19	北方

图表 4-3-4 发言人信息表示例

录音编号：每个发言人领取 1 个录音编号，每个录音编号对应 1 份录音文本。每个发言人只能参加一次录制。

性别：男 237 人，女 213 人。

年龄：发言人年龄在 13~65 岁。

区域：发言人的籍贯以北方为主，占比 96%。

7 版权声明

本文内容禁止转载，AISHELL(北京希尔贝壳科技有限公司)对本文拥有修改权、更新权及最终解释权。



Copyright