中文普通话语音数据库

AISHELL-ASR0009

产品说明书

北京希尔贝壳科技有限公司

bd@aishelldata.com

目录

1 产品概述.....2
2 录音语料.....2
2.1 语料池的制作.....2
2.1.1 语料池内容.....2
2.1.2 语料池处理.....3
2.2 录音文本的结构设计.....3
3 发音人信息.....4
3.1 基本信息记录.....4
3.2 发音人结构特征.....4
3.2.1 性别比例.....4
3.2.2 年龄比例.....4
3.2.3 方言区域比例.....5
4 数据录制环境.....5
4.1 录制环境.....5
4.2 录制设备.....5
4.3 录制方法.....5
5 标注转写规范.....6
6 产品目录结构.....7
6.1 目录结构.....7
6.2 命名规则.....7
7 版权声明.....8

北京希尔贝壳科技有限公司
bd@aishelldata.com

1 产品概述

AISHELL-ASR0009 中文普通话语音数据库共 3002 小时。录音文本涉及智能家居、无人驾驶、工业生产等 15 个领域。邀请 2000 名来自中国不同口音区域发音人参与录制。录制过程在安静室内环境中，同时使用 3 种不同设备：高保真麦克风（44.1kHz，16bit，999.7H）；Android 系统手机（16kHz，16bit，1000.7H）；iOS 系统手机（16kHz，16bit，1002.4H）。

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 97%以上。

2 录音语料

2.1 语料池的制作

2.1.1 语料池内容

考虑到语音识别在智能家居、无人驾驶、工业生产等领域的应用，语料在 15 个领域中选定，共 300 万句常用中文语句。

C0001-C1000		C9001-C9348	
序号	**领域**	**序号**	**领域**
1	家居家电名称和控制命令	1	家居
2	POI（地理信息）	2	POI
3	音乐类（语控）	3	音乐类（语控）
4	数字类（语控）	4	数字类（语控）
5	电视、电影名称	5	电视名称
6	财经	6	财经-政策
7	科技	7	财经-地产
8	体育	8	科技-国际
9	娱乐	9	科技-国内
10	时事新闻	10	体育
11	英文拼读	11	娱乐-欧美
		12	娱乐-港台

北京希尔贝壳科技有限公司

bd@aishelldata.com

		13	时事新闻
		14	综合
		15	英文拼读

图表 2-1 语料池内容

2.1.2 语料池处理

- 脱敏处理。删除政治敏感、个人隐私、色情暴力等内容。
- 删除 <,>,[,], ～ ,/,\,= 等符号。
- 删除含有中文和英文以外语言的内容。
- 统一格式。

2.2 录音文本的结构设计

考虑到语音覆盖及音素平衡，此数据库 C0001-C1000 录音文本采用每份 500 句的分配方式设计，C9001-C9348 录音文本采用每份 520 句的分配方式设计，从语料池中抽取，结构如下。

C0001-C1000			C9001-C9348		
序号	**领域**	**每份分配量/句**	**序号**	**领域**	**每份分配量/句**
1	家居家电和控制名称	5	1	家居	5
2	POI（地理信息 ）	30	2	POI	30
3	音乐类（ 语控）	46	3	音乐类（语控）	46
4	数字类（语控）	29	4	数字类（语控）	29
5	电视、电影名称	10	5	电视名称	13
6	财经	132	6	财经-政策	20
7	科技	85	7	财经-地产	20
8	体育	66	8	科技-国际	15
9	娱乐	27	9	科技-国内	10
10	时事新闻	66	10	体育	20
11	英文拼读	4	11	娱乐-欧美	4
			12	娱乐-港台	10
			13	时事新闻	10
			14	综合	200
			15	英文拼读	4
			16	问答	84
合计	11 项	500 句	合计	16 项	520 句

图表 2-2 录音文本结构

北京希尔贝壳科技有限公司

bd@aishelldata.com

3 发音人信息

3.1 基本信息记录

发音人信息记录内容包括任务编号、性别、年龄段、籍贯。

任务编号	年龄区间	性别	口音区域
C0001	B	女	北方

图表 3-1 发言人信息表示例

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

年龄区间：A(18 岁以下)、B(18-25 岁)、C(26-40 岁)、D(41 岁以上)。

性别：男-男性；女-女性。

口音区域：按照发言人原生语言所属区域，分为北方、南方、其他。

3.2 发音人结构特征

3.2.1 性别比例

数据库总人数为 2000 人，男 863 人，女 1137 人。

性别	男性	女性	合计
比例	43%	57%	100%

图表 3-2-1 性别分布

3.2.2 年龄比例

A(18 岁以下)18 人、B(18-25 岁)1528 人、C(26-40 岁)310 人、D(41 岁以上)144 人。

年龄段	人数	比例	男性	女性
A	18	9%	5	13
B	1528	76.4%	633	895
C	310	15.5%	159	151
D	144	7.2%	66	78
合计	**2000**	**100.00%**	**863**	**1137**

北京希尔贝壳科技有限公司

bd@aishelldata.com

图表 3-2-2 年龄分布

3.2.3 方言区域比例

方言区域	人数	比例
北方	1267	63%
南方	710	36%
其他	23	1%

图表 3-2-3 方言区域分布

4 数据录制环境

4.1 录制环境

安静室内，不包括明显的其他人说话声音及其他噪音，无回音。发言人按照正常语速，朗读录音文本。

4.2 录制设备

录制设备包括高保真麦克风和录音机、手机。本数据库数据存储格式为高保真录制数据 44.1kHz、16bit 单声道和手机录制数据 16kHz、16bit 单声道两种格式。

4.3 录制方法

发音人距离高保真麦克风 20 厘米，以讲话正常音量，正常语速，朗读录音文本。Android 系统手机与 iOS 系统手机分别与麦克风间隔 20 厘米布置。

图表 4-3 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，“N B A”、“U F O”。

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。

北京希尔贝壳科技有限公司

bd@aishelldata.com

6 产品目录结构

6.1 目录结构

数据目录结构	
数据目录结构	
| AISHELL-ASR0009.pdf	**(数据库简介)**
└─DOC	**(文本说明文件)**
||├─wav_list.txt	**(音频列表)**
||├─content.txt	**(语音文本内容)**
||├─spk_info.xlsx	**(录音人信息)**
|	
└─SPEECHDATA	**(数据文件夹)**
|├─C0001	**(录音人文件夹)**
|| || ANDROID	**(设备文件夹)**
|| || AC0001W0001.wav	**(音频文件)**
|| || AC0001W0001.txt	**(语音内容文本)**
|| || IOS	**(设备文件夹)**
|| || IC0001W0001.wav	**(音频文件)**
|| || IC0001W0001.txt	**(语音内容文本)**
|| || MIC	**(设备文件夹)**
|| || MC0001W0001.wav	**(音频文件)**
|| || MC0001W0001.txt	**(语音内容文本)**

图表 6-1-1 数据目录结构

6.2 命名规则

CORPUS/USAGE/SPEAKER_NUM/EQUIPMENT_ID/SPEECH_ID

e.g. AISHELL-ASR0009/SPEECHDATA/C0001/ANDROID/ H0001A0001.wav

目录名称	内容	备注
CORPUS	AISHELL-ASR0009	语音数据库编号
USAGE	SPEECHDATA	文件夹名称
SPEAKER_NUM	C0001	录音人文件夹
EQUIPMENT_ID	ANDROID/IOS/ MIC	设备文件夹
SPEECH_ID	AC0001W0001.wav	WAV 文件

图表 6-2-1 命名规则

7 版权声明

本文章内容禁止转载，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。

北京希尔贝壳科技有限公司

bd@aishelldata.com