粤语方言语音数据库
AISHELL-ASR0068
产品说明书

目录

北京希尔贝壳有限公司

bd@aishelldata.com

1 产品概述

AISHELL-ASR0068 粤语方言语音数据库共 768 小时。数据类型包含朗读和对话，其中朗读 618.8 小时，对话 149.4 小时。录音语言为粤语；录音地区为中国广州。共邀请 2672 名录音人，在安静环境下使用手机进行录制。

此数据库经过专业校对，文本正确率达到 95%以上，可用于声纹识别、语音识别等研究。

2 录音语料

2.1 录音文本结果

此数据库朗读部分录音文本采用每人 470 句，语音内容覆盖唤醒词、命令词、自由文本等，另一批数据主要录制车载语控部分，每人约 10~15 分钟，具体分配如下：

PY0001-PY3400			PY3501-PY5128	
领域	内容例句	编号	领域	
唤醒词	小爱同学、小度小度	0001-0021	车载语控	0001-0300
POI	导航到科技广场	0022-0044		
家居控制	请帮我开启视频二	0045-0051		
拼读	Q M P E	0052-0054		
数字	八百六十零九千五百七十七	0055-0074		
问答	陈十三演的电影有哪些	0075-0089		
音乐	唱首歌汪苏泷的歌	0090-0119		
电视电影	游戏女王国语版	0120-0139		
电台	我要把 A M 调为九百一十五点五千兆	0140-0144		
娱乐	欧阳娜娜之后到大陆工作	0145-0160		
财经	同比增长百分之幺点九	0161-0240		
科技	出货量总额不敌苹果手表	0241-0290		
体育	今后阿拉会以北马为基础	0291-0340		
时事新闻	玻璃门上贴了一张告示	0341-0390		
综合	对你来说今晚一定很漫长	0391-0470		
合计		470		300

图表 2-1-1

车载语控部分针对车控、音乐、导航、问询等类别句数占比如下：

北京希尔贝壳有限公司

bd@aishelldata.com

领域	占比
音乐	50.0%
POI	27.4%
问询	3.8%
车控	7.8%
其它	11.0%
合计	100%

图表 2-1-2

对话时录音人根据给定主题，自由发挥进行录制，主题分类大致如下：

领域	占比
电商零售	14%
金融	12%
教育	13%
科技	7%
生活服务	13%
汽车销售	7%
房地产	9%
物流	5%
其它	20%

图表 2-1-3

2.2 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，数字串部分除外。
- 删除含有中文和英文以外语言的内容。
- 统一格式。

3 发音人信息

3.1 性别比例

数据库总人数为 2672 人，男 1125，女 1547 人。（图表 3-1）

性别	男性	女性	合计
占比	42.1%	57.9%	100%

图表 3-1

3.2 年龄比例

录音人年龄具体分布如下所示：A(18-25 岁)；B(26-40 岁)；C(41-50 岁)；D(51+岁)。（图表 3-2）

	年龄段	人数	占比
A	18-25 岁	686	25.6%
B	26-40 岁	1708	63.9%
C	41-50 岁	223	8.4%
D	51+岁	55	2.1%
合计		2672	100%

图表 3-2

4 数据录制环境

4.1 录制环境

安静室内，无回音，无其它人说话。

4.2 录制设备

市面上常见的手机，如苹果、华为、小米、oppo 等。数据存储格式为 16kHz，16bit wav 文件。

4.3 录制方法

朗读部分录音人在安静室内通过手机小程序进行录制，正常语速朗读文本。

对话部分录音人在安静室内通过手机拨打电话进行录制。

5 语音数据转写

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。一般准则如下：

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）音频中有英文发音的应写成相应的汉字或英文。具体分为以下几种情况：

网址中包含的所有的字母或单词，均为大写。例如：发音内容为”www.abc.com”，应转写为“三 W 点 A B C 点 com”

发音中包含的英文单词，转写时全部为小写。

发音中包含的英文字母，转写时全部为大写。

对于一些大写专有名词，大写且不加空格，例如：CEO、NBA 等。

4）标注内容的完整性要与实际发音一致，不得删减。

6 数据文件目录

6.1 目录结构

数据目录树	
数据目录结构	
| AISHELL-ASR0068.pdf	**(数据库简介)**
└─dialogue	**(对话数据)**
|├─DOC	**(文本文件夹)**
|├─wav_list.txt	**(音频列表)**
|├─spkrinfo.xlsx	**(录音人信息)**
|├─Textgrid	**(转写文本文件夹)**
|├─PYD_0623.Textgrid	**(转写文本)**
|├─SPEECHDATA	**(数据文件夹)**
|| || PYD_0623.wav	**(音频文件)**
|	
└─reading	**(朗读数据)**

北京希尔贝壳有限公司

bd@aishelldata.com

```
|├─DOC                                    (文本文件夹)
|  |├─content.txt                         (转写文本)
|  |├─wav_list.xlsx                       (音频列表)
|  |├─spkrinfo.xlsx                       (录音人信息)
|├─SPEECHDATA                             (数据文件夹)
|  |├─PY0028                              (录音人编号)
|  ||  ||  PY0028.wav                     (音频文件)
```

6.2 命名规则

6.2.1 目录命名规则

<DATA_TYPE>/<USAGE>/<SpeakerID>/<wav_ID>

e.g.reading/SPEECHDATA/PY0028/PY0028_0001.wav

目录	内容	备注
DATA_TYPE	dialogue/reading	数据类别
USAGE	SPEECHDATA	文件夹名称
SpeakerID	PY0028	录音人编号
wav_ID	PY0028_0001.wav	音频文件

图表 6-2-1

6.2.2 文件命名规则

<SpeakerID>_<wav_ID>

e.g. PY0028_0001.wav

文件	内容	备注
SpeakerID	PY0028	录音人编号
<SpeakerID>_<wav_ID>	0001.wav	音频 ID

图表 5-2-2

7 版权声明

Copyright

Copyright

北京希尔贝壳有限公司

bd@aishelldata.com