陕西方言语音数据库

AISHELL-ASR0065

产品说明书

北京希尔贝壳有限公司

bd@aishelldata.com

目录

北京希尔贝壳有限公司

bd@aishelldata.com

1 产品概述

AISHELL-ASR0065 陕西方言语音数据库共 738 小时。录音语言为陕西方言；录音地区为中国。共邀请 8356 名录音人，在安静环境下使用手机电话呼叫基站录音。

数据经过专业校准，文本正确率达到 97%以上，此数据库可用于声纹识别、语音识别等研究。

2 发音人信息

2.1 性别比例

数据库总人数为 8356 人，男 4086 人，女 4270 人。（图表 2-1）

性别	男性	女性	合计
比例	48.9%	51.1%	100%

图表 2-1

2.2 年龄比例

录音人年龄覆盖 20-66 岁，具体分布如下所示：A(20-30 岁)；B(31-40 岁)；C(41-50 岁)；D(51-60 岁)。（图表 2-2）

	年龄段	人数	比例
A	20-30 岁	1763	21.10%
B	31-40 岁	3304	39.54%
C	41-50 岁	2474	29.61%
D	51-60 岁	815	9.75%
合计		8356	100%

图表 2-2

北京希尔贝壳有限公司

bd@aishelldata.com

2.3 地域比例

录音人来自陕西，其中关中、陕北、陕南分布如下所示：

地域	人数	占比
关中	4996	59.8%
陕南	1639	19.6%
陕北	1721	20.6%
合计	8356	100%

图表 2-3

3 数据录制环境

3.1 录制环境

安静室内，无回音，无其它人说话。

3.2 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 8kHz，16bit wav 文件。

3.3 录制方法

录音人在安静室内手持手机呼出电话到基站，正常语速朗读文本。

4 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。

5 数据文件目录

5.1 目录结构

数据目录树	
数据目录结构	
| AISHELL-ASR0065.pdf	**(数据库简介)**
└─DOC	**(文本说明文件)**
||├─content.txt	**(转写文本)**
||├─wav_list.txt	**(音频列表)**
||├─spkrinfo.xlsx	**(录音人信息)**
|	
└─SPEECHDATA	**(数据文件夹)**
|├─S0457	**(录音人编号)**
|| || S0457_0001.wav	**(音频文件)**

5.2 命名规则

5.2.1 目录命名规则

<USAGE>/<SpeakerID>/<wav_ID>

e.g. SPEECHDATA/S0457/S0457_0001.wav

目录	内容	备注
USAGE	SPEECHDATA	文件夹名称
SpeakerID	S0457	录音人编号
wav_ID	S0457_0001.wav	音频文件

图表 5-2-1

5.2.2 文件命名规则

<SpeakerID>_<wav_ID>

e.g. S0457_0001.wav

文件	内容	备注
SpeakerID	S0457	录音人编号
<SpeakerID>_<wav_ID>	0001.wav	音频 ID

图表 5-2-2

6 版权声明