

上海方言语音数据库 AISHELL-ASR0066



产品说明书
希尔贝壳
A I S H E L L

Copyright

目录

1 产品概述	2
--------------	---

2 录音语料	2
2.1 录音文本结果	2
2.2 语料池处理	3
3 发音人信息	3
3.1 性别比例	3
3.2 年龄比例	3
4 数据录制环境	4
4.1 录制环境	4
4.2 录制设备	4
4.3 录制方法	4
5 语音数据转写	4
6 数据文件目录	5
6.1 目录结构	5
6.2 命名规则	6
6.2.1 目录命名规则	6
6.2.2 文件命名规则	6
7 版权声明	6

1 产品概述

AISHELL-ASR0066 上海方言语音数据库共 340 小时。数据类型包含朗读和对话，其中朗读 174 小时，对话 166 小时。录音语言为上海方言；录音地区为中国上海。共邀请 724 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，文本准确率达到 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，具体分配如下：

领域	内容例句	编号
唤醒词	小爱同学、小度小度	0001-0021
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	同比增长百分之么点九	0161-0240
科技	出货量总额不敌苹果手表第一天额预定量	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

图表 2-1-1

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	12%
金融	14%
教育	13%
科技	13%
生活服务	13%
汽车销售	7%
房地产	9%
物流	6%
其它	15%

图表 2-1-2

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

Copyright

3.1 性别比例

数据库总人数为 724 人，男 232 人，女 492 人。（图表 3-1）

性别	男性	女性	合计
占比	32%	68%	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	79	10.9%
B	26-40 岁	374	51.7%
C	41-50 岁	223	30.8%
D	51+岁	48	6.6%
合计		724	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。

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-ASR0066.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)
├─PHD_0006.Textgrid	(转写文本)
├─SPEECHDATA	(数据文件夹)
PHD_0006.wav	(音频文件)
└─reading	(朗读数据)
├─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)
├─PH0012	(录音人编号)
PH0012_0001.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/PH0012/PH0012_0001.wav

目录	内容	备注
DATA_TYPE	dialogue/reading	数据类别
USAGE	SPEECHDATA	文件夹名称
SpeakerID	PH0012	录音人编号
wav_ID	PH0012_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. PH0012_0001.wav

文件	内容	备注
SpeakerID	PH0012	录音人编号
<SpeakerID>_<wav_ID>	0001.wav	音频 ID

图表 6-2-2

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