中文普通话语音数据库

数据产品说明书

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail： bd@aishelldata.com

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com

1 产品概述

此中文普通话语音数据库共 1000 小时。录音文本涉及智能家居、无人驾驶、工业生产等 11 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，1000H）；Android 系统手机（16kHz，16bit，1000H）；iOS 系统手机（16kHz，16bit，1000H）。

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 95%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 11 个领域中选定，共 50 万句常用中文语句。（图表 2-1）

序号	领域
1	家居家电名称和控制命令
2	POI（地理信息）
3	音乐类（语控）
4	数字类（语控）
5	电视、电影名称
6	财经
7	科技
8	体育
9	娱乐
10	时事新闻
11	英文拼读

图表 2-1 语料池内容

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com

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。
- 删除单句含有 25 字以上的内容。
- 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，此数据库录音文本采用每份 500 句的分配方式设计，从语料池中抽取，结构如下。（图表 2-2）

序号	领域	每份分配量/句
1	家居家电和控制名称	5
2	POI（地理信息）	30
3	音乐类（语控）	46
4	数字类（语控）	29
5	电视、电影名称	10
6	财经	132
7	科技	85
8	体育	66
9	娱乐	27
10	时事新闻	66
11	英文拼读	4
合计	**11 项**	**500 句**

图表 2-2

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、口音区域、年龄区间、籍贯。（图表 3-1）

任务编号	性别	口音区域	年龄区间	籍贯
0001	男	北方	A	河北

图表 3-1

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com

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。

口音区域：按照发言人原生语言所属区域，分为北方、南方、粤贵闽、其他。

年龄区间：A(16-25 岁)、B(26-40 岁)、C(41 岁以上)。

籍贯：记录发言人身份证所示籍贯信息。

3.2 发音人结构特征

3.2.1 性别比例

性别	男性	女性	合计
比例	47%	53%	100%

图表 3-2-1

3.2.2 年龄比例

A

	年龄段	比例
A	16-25 岁	79%
B	26-40 岁	18%
C	> 41 岁	3%
合计		100%

图表 3-2-2

3.2.3 方言区域比例

区域	比例
北方	83%
南方	10%
粤贵闽	4%
其他	3%
合计	100%

图表 3-2-3

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com

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、iOS 系统手机、Android 系统手机。

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）

图表 4-3

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：

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com

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，或者一些英文缩写全部大写加空格，例如：C E O、C C T V 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-ASR0009-[ZH-CN]_中文普通话语音数据_产品说明书.docx	**（数据库简介）**
└─DOC	**（文本说明文件）**
||├─all_wav_list.txt	**（音频列表）**
||├─content.txt	**（转写内容列表）**
||├─readme.txt	**（目录说明文件）**
||├─spkrinfo.xlsx	**（录音人信息）**
|	
└─SPEECHDATA	**（数据文件夹）**
||├─C0001	**（录音人文件夹）**
|| || MIC	**（高保真麦克风数据）**
|| || MC0001W0001.wav	**（音频文件）**
|| || MC0001W0001.txt	**（语音内容文本）**
|| || IOS	**（iOS 系统手机数据）**
|| || IC0001W0001.wav	**（音频文件）**
|| || IC0001W0001.txt	**（语音内容文本）**
|| || ANDROID	**（android 系统手机数据）**
|| || AC0001W0001.wav	**（音频文件）**
|| || AC0001W0001.txt	**（语音内容文本）**

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com

6.2 命名规则

6.2.1 目录命名规则 （图表 6-2-1）

/<CORPUS>/<USAGE>/<FILE_ID>/<SPEECH_ID>

e.g.AISHELL-ASR0009-[ZH-CN]/SPEECHDATA/C0001/MIC/MC0001W0001.wav

目录	内容	备注
CORPUS	AISHELL-ASR0009	数据库名称编号
USAGE	SPEECHDATA	数据存放文件夹名称
RECORDER_ID	MIC/IOS/ANDROID	录制设备分类文件夹
FILE_ID	C0001	录音人文件夹名称
SENTENCE_ID	MC0001W0001.txt	TXT 文件
SPEECH_ID	MC0001W0001.wav	WAV 文件

图表 6-2-1

6.2.2 文件命名规则 （图表 6-2-2）

<RECORDER_ID><SPEAKER_IC><WAV_NUM>

e.g. MC0001W0001.wav

文件	内容	备注
RECORDER_ID	M/I/A	录制设备分类文件夹
SPEAKER_NUM	C0001	录音人 ID
WAV_NUM	W0001	WAV 编号

图表 6-2-2

7 版权声明

北京希尔贝壳科技有限公司
Beijing Shell Shell Technology Co.,Ltd
Add: Room 3-621, 6F, Zhongguancun Lifangting No. 1, Shanyuan Road, Haidian District, Beijing 100080, P.R.China
Tel：+86 10 80225006 E-mail：bd@aishelldata.com