

印地语语音数据库

[AISHELL-ASR0026]

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

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.....	2
2.2 录音文本的结构设计.....	3
3 发音人信息.....	3
3.1 基本信息记录.....	3
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
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 产品目录结构.....	5
5.1 目录结构.....	5
5.2 命名规则.....	5
6 版权声明.....	5

1 产品概述

AISHELL-ASR0026 语音数据库共 550 小时。录音语言, 印地语。录音地区, 印度。录音文本涉及教育、科技、娱乐、健康等 10 个领域。580 名来自印度不同区域的发言人参与录制。录制过程在安静室内环境中, 使用手机 (16kHz, 16bit) 进行录制。

经过专业语音校对人员转写标注, 并通过严格质量检验, 此数据库文本正确率在 95% 以上。

2 录音语料

2.1 语料池的制作

2.1.1 语料池内容

AISHELL-ASR0026 录音文本领域统计表。

序号	领域
1	教育
2	娱乐
3	科技
4	健康
5	生活方式&时尚
6	新闻
7	社交闲聊
8	商业&金融
9	体育
10	小说

图表 2-1-1 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，此数据库录音文本采用每份 420 句的分配方式设计，从语料池中抽取，结构如下。

序号	领域	每人分配/句
1	教育	40
2	娱乐	30
3	科技	35
4	健康	20
5	生活方式&时尚	50
6	新闻	80
7	社交闲聊	120
8	商业&金融	20
9	体育	20
10	小说	5
合计		420

图表 2-2-2 录音文本结构设计

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、年龄区间。

任务编号	性别	年龄区间
H0001	男	A

图表 3-1-1 发言人信息示例

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(40 岁以上)。

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	53%	47%	100%

图表 3-2-1

3.2.2 年龄比例

	年龄段	比例
A	< 18 岁	5%
B	18-25 岁	68%
C	26-40 岁	19%
D	> 41 岁	9%
合计		100%

图表 3-2-2

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 iOS 系统手机、Android 系统手机。

4.3 录制方法

发音人距离手机 20 厘米, 以讲话正常音量, 正常语速, 朗读录音文本。

5 产品目录结构

5.1 目录结构

数据目录结构	
数据目录结构	
AISHELL-ASR0026.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)
├─H0005	(录音人文件夹)
H0005W0001.wav	(音频文件)
H0005W0001.txt	(文本文件)

图表 5-1-1 数据目录结构

5.2 命名规则

CORPUS/USAGE/SPEAKER_NUM/SPEECH_ID

e. g. AISHELL-ASR0026/SPEECHDATA/H0005/H0005W0001. wav

目录名称	内容	备注
CORPUS	AISHELL-ASR0026	语音数据库编号
USAGE	SPEECHDATA	文件夹名称
SPEAKER_NUM	H0005	录音人文件夹名称
SPEECH_ID	H0005W0001.wav	WAV 文件
SENTENCE_ID	H0005W0001.txt	TXT 文件

图表 5-2-1 命名规则

6 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