粤语语音数据库

AISHELL-ASR0079

产品说明书

北京希尔贝壳科技有限公司
bd@aishelldata.com

目录

1 产品概述

AISHELL-ASR0079 粤语语音数据库，共 965 小时。采样率 16000Hz，比特率 16bit。 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 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）数字要转写为英文拉丁形式, 比如“一二三”，而不是“123”。

3）英文按单词小写，拼读发音需写成大写字母加空格，如，“W W W dot Google dot com”，“C E O”。

4）标注内容的完整性要与实际发音一致，不得删减。

3 目录结构与命名规则

3.1 目录结构

数据目录树	
数据目录结构	
| AISHELL-ASR0079.pdf	**(数据库简介)**
└─DOC	**(文本说明文件)**
||├─content.txt	**(转写内容列表)**
└─WAV	**(音频文件夹)**
||├─Y0000001	**(音频文件)**

图表 3-1-1 目录结构

3.2 命名规则

CORPUS/USAGE/SPEECH_ID

e.g.AISHELL-ASR0079/WAV/Y0000001.wav

目录名称	内容	备注
CORPUS	AISHELL-ASR0079	语音数据库编号
USAGE	WAV	音频文件夹
SPEECH_ID	Y0000001.wav	音频文件

图表 3-2-1 命名规则

4 版权声明