

语音数据库说明书

AISHELL-ASR0056



希尔贝壳
A I S H E L L

Copyright

目录

1 产品概述.....	3
2 产品目录结构.....	3
2.1 目录结构.....	3
2.2 命名规则.....	3
3 标注转写规范.....	3
4 版权声明.....	4



Copyright

1 产品概述

AISHELL-ASR0056 语音数据库共 1200 小时。录音语言，中文、英文；录音地区，中国。

采样率 16Khz，比特率 16bit。

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 97%以上。

2 产品目录结构

2.1 目录结构

数据目录结构	
数据目录结构	
AISHELL-ASR0056.pdf	(数据库简介)
└─SPEECHDATA	(数据文件夹)
─CE01	(block 文件夹)
─CE01000002.wav	(音频文件)
└─DOC	(文本文件夹)
─wav_list.txt	(音频列表)
─content.txt	(转写结果)

图表 2-1-1 目录结构

2.2 命名规则

AISHELL-ASR0056\SPEECHDATA\wav\<blockID>\<wavID>.wav

e.g. AISHELL-ASR0056\SPEECHDATA\wav\CE01\CE01000002.wav

名称	内容	备注
<blockID>	CE01-CE20 & PCE01-PCE10	block 编号

图表 2-2-1 路径命名规则

3 标注转写规范

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。

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) 句中出现的英文按照发音写出单词，如“thank you”。按拼读朗读的字母，需转写成大写字母加空格的形式。如，“NBA”、“UFO”。
- 4) 句中包含的符号，按实际发言人发音转写。如“三 W 点 百度 点 com”。没有发音的符号，需要删掉。品牌名称，专有名称等按照实际惯用格式转写，如“QQ 空间”、“iPhone”、“喜马拉雅”。
- 5) 标注内容的完整性要与实际发音一致，不得删减。
- 6) 标注数据覆盖性别以及环境信息。

4 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