# 台湾普通话语音数据库

## AISHELL-ASR0082

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-ASR0082 台湾普通话语音数据库，共 720.8 小时。采样率 16000Hz，比特率 16bit。数据为网络数据，包含健康、影视、旅游、科技等 11 个领域，该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 96%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）数字要转写为汉字形式，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-ASR0082.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)** |
| **\|├─健康** | **(音频领域)** |
| **\|\|├─tw9000101_1_4600.wav** | **(音频文件)** |

图表 3-1-1 目录结构

## 3.2 命名规则

CORPUS/USAGE/DOMAIN/SPEECH_ID

e.g.AISHELL-ASR0082/SPEECHDATA/健康/tw9000101_1_4600.wav

| 目录名称 | 内容 | 备注 |
|---|---|---|
| CORPUS | AISHELL-ASR0082 | 语音数据库编号 |
| USAGE | SPEECHDATA | 音频文件夹 |
| DOMAIN | 健康 | 音频领域 |
| SPEECH_ID | U0000001.wav | 音频文件 |

图表 3-2-1 命名规则

# 4 领域划分

该数据库总共划分为 11 个领域，包含健康、经济、影视、科技等，具体占比如下所示：

| 领域 | 时长 | 占比（%） |
|---|---|---|
| 健康 | 18.11 | 2.51 |
| 影视 | 166.47 | 23.1 |
| 旅游 | 10.46 | 1.45 |
| 时尚 | 59.46 | 8.25 |
| 时政 | 102.96 | 14.28 |
| 星座 | 95.78 | 13.29 |
| 游戏 | 46.88 | 6.5 |
| 科技 | 62.38 | 8.65 |
| 经济 | 12.28 | 1.7 |
| 综艺 | 62.73 | 8.7 |
| 饮食 | 83.36 | 11.57 |
| **总计** | **720.8** | **100** |

图表 4-1 领域划分



# 5 版权声明