

# 会议演讲音频数据库

[AISHELL-ASR0037]



希尔贝壳  
A I S H E L L

产品说明书

Copyright

北京希尔贝壳科技有限公司

Beijing Shell Shell Technology Co.,Ltd

Add: Room 813, Building No.4 Shangdi 10th Street, Haidian District, Beijing 10080, 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 .....            | 3 |
| 2.2 录音文本的结构设计.....           | 3 |
| 3 语音数据转写.....                | 3 |
| 4 数据文件目录.....                | 4 |
| 4.1 目录结构.....                | 4 |
| 4.2 命名规则.....                | 4 |
| 4.2.1 目录命名规则（图表 4-2-1） ..... | 4 |
| 4.2.2 文件命名规则（图表 4-2-2） ..... | 5 |
| 5 版权声明.....                  | 5 |

# 1 产品概述

会议演讲音频数据库 AISHELL-ASR0037 录音时长 1459 小时，音频来自互联网会议，为真实会场的中控台录音，音频采样率为 16kHz，16bit。录音文本涉及科技、金融、教育等 16 个领域。

数据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 97% 以上。

## 2 录音语料

### 2.1 语料池的制作

#### 2.1.1 语料池内容

考虑到语料多样性，语料在 17 个领域中选定。

| 序号 | 领域 ID | 内容     |
|----|-------|--------|
| 1  | MTGC  | 科技     |
| 2  | MTGA  | 金融     |
| 3  | MTGD  | 医疗健康   |
| 4  | MTGE  | 体育     |
| 5  | MTGF  | 教育     |
| 6  | MTGG  | 娱乐     |
| 7  | MTGI  | 房地产    |
| 8  | MTGH  | 文学艺术   |
| 9  | MTGJ  | 交通物流   |
| 10 | MTGM  | 政府报告   |
| 11 | MTGK  | 汽车     |
| 12 | MTGB  | 生活服务   |
| 13 | MTGO  | 创业管理   |
| 14 | MTGL  | 电商零售   |
| 15 | MTGN  | 人物访谈演讲 |
| 16 | MTGP  | 多媒体    |
| 17 | MTGQ  | 其他     |

图表 2-1 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。
- 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，此数据库录音文本领域覆盖 17 类，具体结构如下。

| 序号 | 领域     | 每份分配量/占比(%) |
|----|--------|-------------|
| 1  | 科技     | 43.78       |
| 2  | 金融     | 9.54        |
| 3  | 医疗健康   | 3.28        |
| 4  | 体育     | 0.72        |
| 5  | 教育     | 3.94        |
| 6  | 娱乐     | 2.55        |
| 7  | 房地产    | 0.54        |
| 8  | 文学艺术   | 1.49        |
| 9  | 交通物流   | 0.24        |
| 10 | 政府报告   | 1.25        |
| 11 | 汽车     | 1.59        |
| 12 | 生活服务   | 2.41        |
| 13 | 创业管理   | 3.31        |
| 14 | 电商零售   | 1.78        |
| 15 | 人物访谈演讲 | 8.81        |
| 16 | 多媒体    | 1.97        |
| 17 | 其他     | 12.8        |
| 合计 |        | <b>100</b>  |

图表 2-2

## 3 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，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：CEO、CCTV 等。

4) 标注内容的完整性要与实际发音一致，不得删减。

## 4 数据文件目录

### 4.1 目录结构

| 数据目录树                       |          |
|-----------------------------|----------|
| 数据目录结构                      |          |
| AISHELL-ASR0037.pdf         | (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) |
|                             |          |
| └─SPEECHDATA                | (数据文件夹)  |
| ├─MTGA                      | (领域文件夹)  |
| ├─├─ MTGA0036A001W00131.wav | (音频文件)   |

### 4.2 命名规则

#### 4.2.1 目录命名规则（图表 4-2-1）

/<USAGE>/<FILE\_ID>/<SPEECH\_ID>

e. g. SPEECHDATA/MTGA/MTGA0036A001W00131. wav

| 目录        | 内容                     | 备注        |
|-----------|------------------------|-----------|
| USAGE     | SPEECHDATA             | 文件夹名称     |
| FILE_ID   | MTGA                   | 文本领域文件夹名称 |
| SPEECH_ID | MTGA0036A001W00131.wav | WAV 文件    |

图表 4-2-1

## 4.2.2 文件命名规则（图表 4-2-2）

<FILE\_ID><SPEAKER\_IC><WAV\_NUM>

e. g. MTGA0036A001W00131.wav

| 文件          | 内容       | 备注      |
|-------------|----------|---------|
| CORPUS_ID   | MTGA0036 | 文本领域编号  |
| SPEAKER_NUM | A0001    | 录音人段落编号 |
| WAV_NUM     | W00131   | WAV 编号  |

图表 4-2-2

## 5 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。



希尔贝壳  
A I S H E L L

Copyright  
Copyright