直播方言语音数据库

AISHELL-ASR0023

产品说明书

目录

1 产品概述

AISHELL-ASR0023 直播方言语音数据库，共 8500 小时。数据主要来自抖音、快手、好看、西瓜视频、哔哩哔哩等视频直播平台的小视频，以及网上公开音视频。方言区覆盖西南官话、粤语、客家等 17 个方言区，每个方言区时长约为 500 小时。采样率 16000Hz，比特率 16bit。总的主播人数超过 3000 人。

2 数据分类

该数据库共覆盖 17 个方言区，每个方言区 500 小时，方言区分类如下：

序号	方言区域	时长
1	粤语	500
2	西南官话	500
3	平话	500
4	晋语	500
5	闽语	500
6	客家话	500
7	吴语	500
8	徽语	500
9	赣语	500
10	湘语	500
11	东北官话	500
12	北京官话	500
13	胶辽官话	500
14	冀鲁官话	500
15	兰银官话	500
16	中原官话	500
17	江淮官话	500
合计		**8500**

数据库的数据 70%来自哔哩哔哩，爱奇艺，抖音，好看，快手，腾讯，西瓜，优酷等常见直播平台的小视频数据，30%来自线上公开的音视频。

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-ASR0023.pdf	**(数据库简介)**
└─DOC	**(文本说明文件)**
||├─wav_list.txt	**(音频列表)**
└─SPEECHDATA	**(音频文件夹)**
||├─BJ	**(方言分类)**
|| || 0001.wav	**(音频文件)**

图表 3-1-1 目录结构

3.2 命名规则

CORPUS/USAGE/SPEAKER_ID/SPEECH_ID

e.g.AISHELL-ASR0023/SPEECHDATA/BJ/0001.wav

目录名称	内容	备注
CORPUS	AISHELL-ASR0023	语音数据库编号
USAGE	SPEECHDATA	音频文件夹
SPEAKER_ID	BJ	方言分类编号
SPEECH_ID	0001.wav	音频文件

图表 3-2-1 命名规则

方言分类编号如下表格所示：

方言分类编号	内容
YY	粤语
XN	西南官话
PH	平话
JY	晋语
MY	闽语
KJ	客家话
WY	吴语
HY	徽语
GY	赣语
XY	湘语
DB	东北官话
BJ	北京官话

JL	胶辽官话
JN	冀鲁官话
LY	兰银官话
ZY	中原官话
JH	江淮官话

图表 3-2-2 方言编号对应表