方言语音数据库

AISHELL-ASR0023-V1

产品说明书

目录

1 产品概述

AISHELL-ASR0023-V1 方言语音数据库，共 169397 小时。数据主要来自抖音、快手、好看、西瓜视频、哔哩哔哩、爱奇艺等视频平台，为网上公开音视频。方言覆盖北京、香港、西藏、四川等 29 个地区。采样率 16000Hz，比特率 16bit，来自 7000 多个博主。

2 数据分类

该数据库共覆盖 26 个地区，方言时长分布分类如下：

序号	方言地区	地区编号	时长	时长占比（100%）
1	安徽	AH	6692	3.95
2	澳门	AM	1461	0.86
3	海南	HAIN	1180	0.70
4	客家	KJ	555	0.33
5	闽南	MN	18364	10.84
6	上海	SH	60	0.04
7	香港	XG	6134	3.62
8	宁夏	NX	2246	1.33
9	江西	JX	3983	2.35
10	山东	SD	10146	5.99
11	河北	HEB	3445	2.03
12	青海	QH	503	0.30
13	河南	HEN	1509	0.89
14	台湾	TW	5200	3.07
15	天津	TJ	8955	5.29
16	湖北	HUB	5002	2.95
17	东北	DB	3440	2.03
18	广东	GD	19487	11.50
19	四川	SC	15002	8.86
20	长沙	CS	649	0.38
21	西藏	XZ	260	0.15
22	新疆	XJ	314	0.19
23	甘肃	GS	2299	1.36
24	重庆	CQ	11790	6.96
25	内蒙	NM	3844	2.27
26	北京	BJ	22792	13.45
27	云南	YN	4507	2.66
28	贵州	GZ	5333	3.15

29	广西	GX	4245	2.51
合计			**169397**	**100**

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-V1.pdf （数据库简介）** **└─SPEECHDATA （音频文件夹）** **｜｜├─BJ （方言分类）** **｜｜ ｜｜ bj0000001.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-V1/SPEECHDATA/BJ/bj0000001.wav

目录名称	内容	备注
CORPUS	AISHELL-ASR0023-V1	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	Bj0000001.wav	音频文件

图表 3-2-1 命名规则