A COUNTRY REPORT — OCOCOSDA ACTIVITIES IN CHINA

Aijun LI , *Dong WANG

Institute of Linguistics, Chinese Academy of Social Sciences
*Research Institute of Information Technology, Tsinghua University

The 24th Conference of the Oriental COCOSDA
18-20 November 2021, Singapore

THE PROJECT FOR THE PROTECTION OF LANGUAGES RESOURCES IN CHINA

Ministry of Education of the People's Republic of China / State Language Commission

Time	The first phase: **2015-2019**. The second phase: **2021-2025**.
Area	**34** provinces, autonomous regions and municipalities across the country.
Language	**123** languages and all Chinese dialects.
Location	**1712**.
Speaker	More than **9000**.
Content	The pronunciation (**1000**), vocabulary (**1200**), sentence (**50**), long-form corpus, oral cultural corpus and all audio and video recordings specified in the survey manual.

The Chinese Language Resource Collection and Service Platform (https://zhongguoyuyan.cn/) **:**

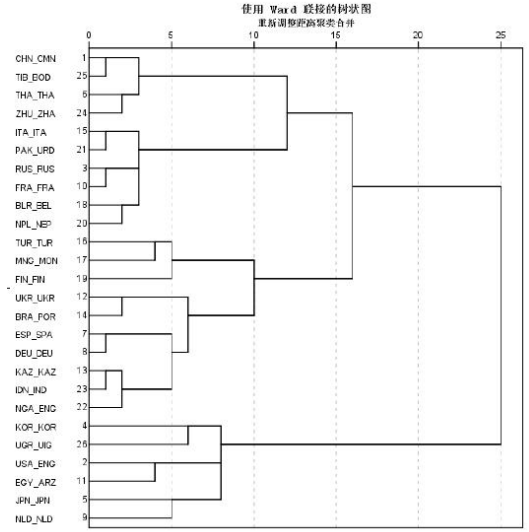
More than **10** million original corpus file data, among them, audio and video data each more than **5** million, with a total physical capacity of **100** TB.

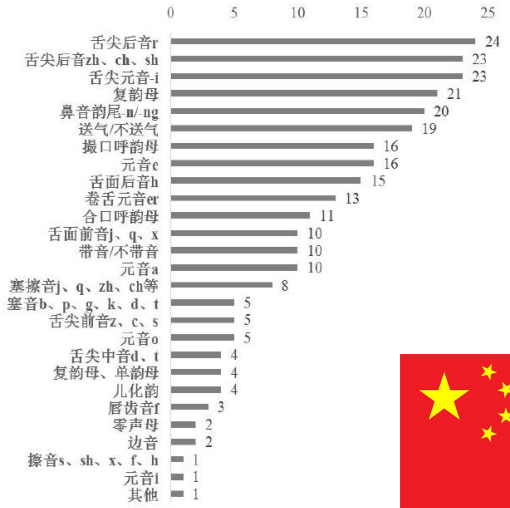
Provided by Center for the Protection and Research of Language Resources of China

GLOBAL ACCENT CHINESE
——The Database of Typical Chinese Speech Errors: From Worldwide Learners (http://www.globalaccentchinese.com/)

- Speech errors collected from Chinese learners in 36 L1s.
- Two basic parts of the database: (1) speech error descriptions (speech error items); (2) audios with typical speech errors from 36 L1s (learners' accent Chinese).
- Accent audios from each L1 consist of sounds of:
- (1) isolate syllables; (2) bi-syllabic/tri-syllabic words; (3) sentences; (4) a dialogue (*a call from sister*, 201 syllables covering 4 tones, 22 initials, 39 finals, neutral tones, all kinds of tone sandhi, interjections and usual sentence types in standard Chinese).

- **4** speakers for each accent Chinese: 2 learners (1 male, 1 female); 2 native speakers (1 male, 1 female).
- ***529*** speech error items from learners ***25*** L1s.
- **≈10010** audio files ((50+30+20+1) *4*25=10100).
- **1** new utility patent.
- **Multiple** explorations have been made based on the database.





NORTHWEST MINZU UNIVERSITY

❖ Ethnic Languages

- Tibetan

- Lexicon : **10k** words
- Text: 100 Mb news, 1GB online text, **100k** Tibetan-Chinese pairs
- Speech: **1000** monosyllabic words produced by **100** speakers from **20** dialects, **30-minute** recordings of **300** speakers

- Mongolian

- Phonetic database: **2000** words produced by **20** speakers
- **200** multimodal phonetic database of Mongolian long tune, Tibetan folk songs, *Hua'er*, *Jianger* and other oral cultures

- Other Ethnic Languages

- Phonetic database of Western Yugu, Eastern Yugu, Dongxiang, Baoan, and Salar: 1000 words produced by 4 speakers each

❖ Gansu Dialects

- **3000** words, **50** sentences, **30-minute** dialogue among three interlocutors, produced by **100** speakers in **27** dialects

❖ Child Language

- **5** speakers of **3-18 months** old

CNCeleb: Multi-Genre Speaker Recognition Corpus

- **3,000** Chinese Celebrities
- **11** copmlex genres
- **803** hours, **659k** utts.
- **Free** for research

Genres	CN-Celeb1			CN-Celeb2		
	# of Spks	# of Utters	# of Hours	# of Spks	# of Utters	# of Hours
Advertisement	17	120	0.18	66	1,542	3.86
Drama	160	7,247	6.43	268	13,116	16.32
Entertainment	483	22,064	33.67	616	31,982	60.84
Interview	780	59,317	135.77	519	34,024	81.28
Live Broadcast	129	8,747	16.35	388	167,019	439.95
Movie	62	2,749	2.20	133	4,449	5.77
Play	69	4,245	4.95	127	14,992	22.04
Recitation	41	2,747	4.98	218	58,231	129.18
Singing	318	12,551	28.83	394	42,157	75.19
Speech	122	8,401	36.22	394	36,680	82.58
Vlog	41	1,894	4.15	488	125,293	177.00
Overall	1,000	130,109	273.73	2,000	529,485	1090.01

Commercial Activities

speechocean 海天瑞声 www.speechocean.com

Languages in China	Speakers: 88,000+ Hours: 78,000+	Mandarin, Accented Mandarin, Cantonese, etc.
English	Speakers: 15,000+ Hours: 15,000+	US English, UK English, etc.
Other Majority Languages	Speakers: 46,000+ Hours: 41,000+	Spanish, Russian, French, German, Italian, Japanese, Korean, etc.
Minority Languages	Speakers: 31,000+ Hours: 37,000+	Afrikaans, Bulgarian, Estonian, Persian, Gujarati , Croatian, Hungarian, Indonesian, Javanese, etc.
TTS Speech Corpus	Speakers: 32 Hours: 450+	Mandarin, Cantonese, English, Spanish, Italian, Portuguese, Japanese, etc.
Lexicon	160+ Languages, 100,000,000+ Entries, including most of the major & minority languages mentioned above.	

MAGIC DATA Data Set Your Mind www.magicdatatech.com

Industry

Financial services | Vehicle | Social | Smart Home | Smart Devices

Datasets

Language	Conversational Speech	Read Speech
Chinese Mandarin Dialect Accented Mandarin Children Mandarin Chinese-English Code-Mixing	78900 Speakers/ 65000 Hours	72600 Speakers/ 25600 Hours
English US English Other Accented English	3000 Speakers/ 12500 Hours	28200 Speakers/ 14300 Hours
Other Languages German,Italian,French,Spanish,Portuguese,Russian Filipino,Indonesian,Malay,Thai,Turkish Japanese,Korean,,Vietnamese,Arabic,Urdu,Hindi	7600 Speakers/ 23500 Hours	16500 Speakers/ 11000 Hours
Total	150,000+ Hours	

希尔贝壳 AISHELL http://www.aishelltech.com

- **Mandarin Corpus**
  - Smart Home
    - 1200 speakers, 3200 Hours
  - Meeting
    - 300 speakers, 2000 Hours
  - TTS
    - 2070 speakers, 700 Hours
  - Other
    - 7000 speakers, 5000 Hours
  - Car EV 2000 speakers, 6000 Hours

- **Hindi Speech Corpus**
  - 600 speakers, 550 Hours
  - 2000 speakers, 6000 Hours
- **Arabic Speech Corpus**
  - 1348 speakers, 899 Hours
  - 2000 speakers, 6000 Hours
- **American English Speech Corpus**
  - 2000 speakers, 987 Hours
  - 2000 speakers, 6000 Hours

Open Source	AISHELL-1	400 speakers, 178H
	AISHELL-2	1991 speakers, 1000H
	AISHELL-3	218 speakers, 85H
	AISHELL-4	60 speakers, 120H
Test Dataset	2018-EVAL Dataset	15 speakers， 5H
	2019-EVAL Dataset	148 speakers， 500H
	2021-EVAL Dataset	3219 speakers， 28H

Challenge Dataset

254 speakers, 1561H
http://www.aishelltech.com/wakeup_data

DMASH 2020
518 speakers， 464H
http://2020.ffsvc.org/

Commercial Activities

DATATANG

www.datatang.com

慧听数据 HUITING DATA

http://www.huitingtech.com/

Chinese	Mandarin/Accented Mandarin Children Mandarin Dialect(Cantonese, Sichuan, Shanghai, Min, Henan, Wuhan, Changsha and etc.) Mongolian/Uyghur/Kazakh/Tibetan Mandarin English Mixed	40000 Hours 90000 Speakers
English	27 countries speaking English, including: US English, UK English, Other Accented English	13000 Hours 26000 Speakers
Other Languages	36 Languages, including: Japanese, Korean, Malay, Indonesian, Russia, French, German, Spanish and etc.	32000 Hours 60000 Speakers
Parallel Corpora	CH-EN, CH-RU, CH-JA, CH-FR, KO-EN, JA-EN and etc.	12 million pairs

Chinese	**Mandarin** **Accented Mandarin** **Children Mandarin** **Elderly Mandarin** **Dialect** **Uygur / Tibetan /Mongolian** **Cantonese** **Mandarin English Mixed** **Cantonese English Mixed**	**40000 Speakers** **30000 Hours**
English	**UK English** **US English** **Other Accented English**	**2200 Speakers** **2100 Hours**
Other Languages **French, German, Italian, Spanish, Mexican Spanish,** **Brazilian Portuguese, Japanese, Korean**		**3900 Speakers** **2100 Hours**

Free Database

Primewords	Mandarin Speech, 1600h	**DataTang**	Mandarin Speech, 1050h	**AIShell**	AI Shell 1-4	**Mobvoi**	Hotwords, 36k utt
MAGICDATA	Mandarin Speech, 755h	**SpeechOcean**	5000 English utt. pronunced by Chinese children. Fluency marked by 5 annotators.			**Databaker**	TTS 10k female, 12h