A COUNTRY REPORT — OCOCOSDA ACTIVITIES IN CHINA

Aijun LI , *Dong WANG

Institute of Linguistics, Chinese Academy of Social Sciences

*Center for Speech and Language Technologies, Tsinghua University

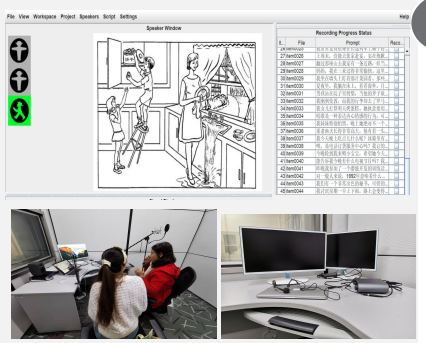
The 25th Conference of the Oriental COCOSDA

24-26 November 2022, Hanoi, Vietnam

SPEECH DATA OF MENTAL DISORDERS ——The Speech Database of Schizophrenia and Their Controls

- A speech database was collected from individuals with schizophrenia by the Shanghai Mental Health Center and their controls by the School of foreign languages at Shanghai Jiao Tong University during 2021-2022.
- This database is one part of the speech database of mental disorders sponsored by the Major Program of National Social Science Foundation of China (No. 18ZDA293)

1



- **Each speaker has two basic parts:**

(1) demographic information, and additional clinical and cognitive information for patients

(2) speech recordings of schizophrenic patients were collected in the hospital, and those of controls were collected in a professional recording room.

2

- **Speech recording lasts for about 45 minutes for each speaker and includes the following three parts:**

Prosody-balanced passage reading | Interview | Picture description

- The recordings were transcribed into texts for the analysis of other language features.
- Part of the findings was reported in the talk "Analysis of the spoken corpus of patients with schizophrenia (精神分裂症患者口语语料分析)" at the 14th PCC in July, Lanzhou.

上海交通大学外国语学院
School of Foreign Languages

Phonetic Distinctive Features of Chinese Dialects

（ http://www.jnuphon.cn/ ）

The Research Institute of Chinese Dialects, Jinan University, Guangzhou, China

Goal

- provide a series of systematic description data of Chinese dialects for people who care about Chinese dialects.
- promote the qualitative and quantitative phonetic research on Chinese Dialects.
- complete the construction of a data base of "The Systematic Distinctive Features on Chinese Dialects ".

辅音类型: 塞音 辅音: 查询

辅音	性别	时常CD(ms)	音强(dB)	VOT(ms)	频率分布		过
					中心频率(hz)	下限频率(hz)	时长(ms)
pa	女	10	1.5	7	2490	2490	15
pa	男	8	-9.0	4	4280	1550	23
pa	男	8	-9.0	4	4280	1550	23
pa	女	10	1.5	7	2490	2490	15
pɐi	男	12	-4.5	8	3620	1420	17
pɐi	女	12	5.0	9	1730	1730	13
pi	男	15	0.5	9	4150	2150	28
pi	女	4	9.0	4	2760	1860	20
piu	男	10	5.0	6	2760	2760	11
piu	女	12	-9.0	8	2950	1920	12
pɛ	男	9	-3.4	5	3610	2350	40
pɛ	女	11	-8.5	9	2710	2710	10

Existing achievements

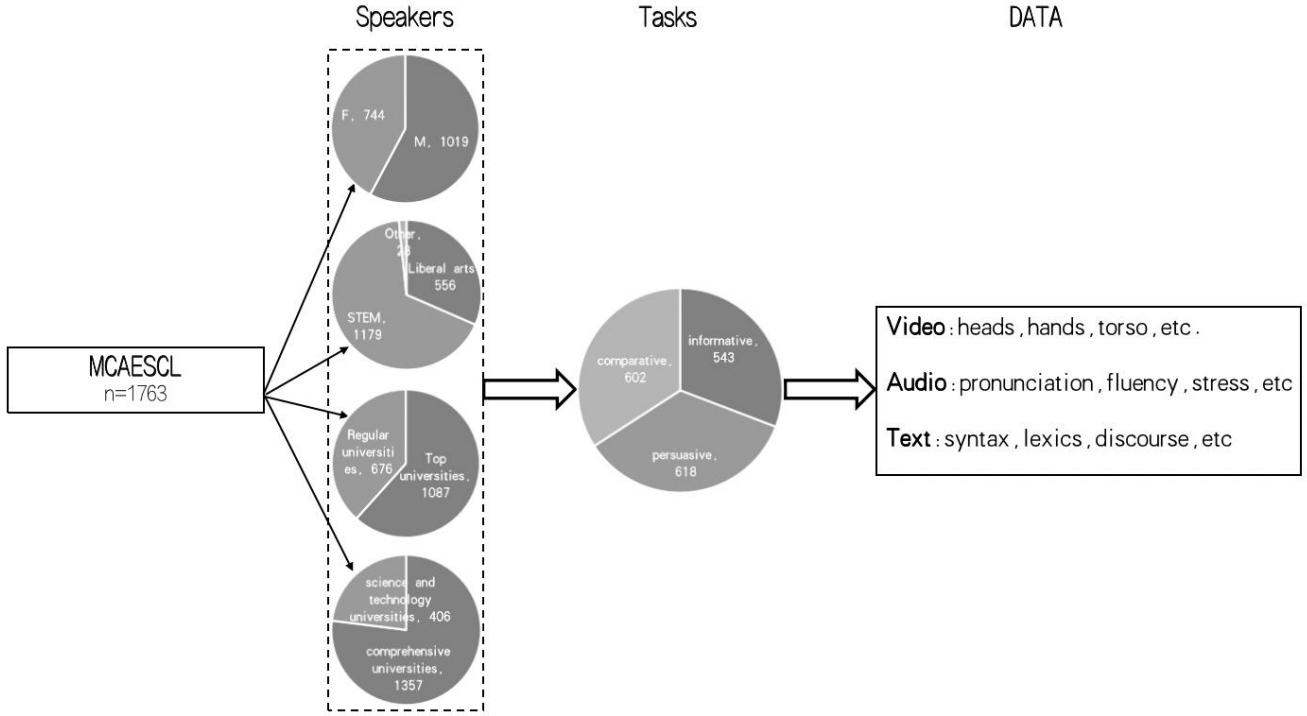
- complete list of Cantonese syllables
- the data of consonants and vowels (monosyllabic and diphthong vowels)
- the pitch data of tones

We welcome all kinds of suggestions for the construction of the website. Your opinions can be expressed through the following methods: website message, email: **tlxzh@jnu.edu.cn**.

MCAESCL

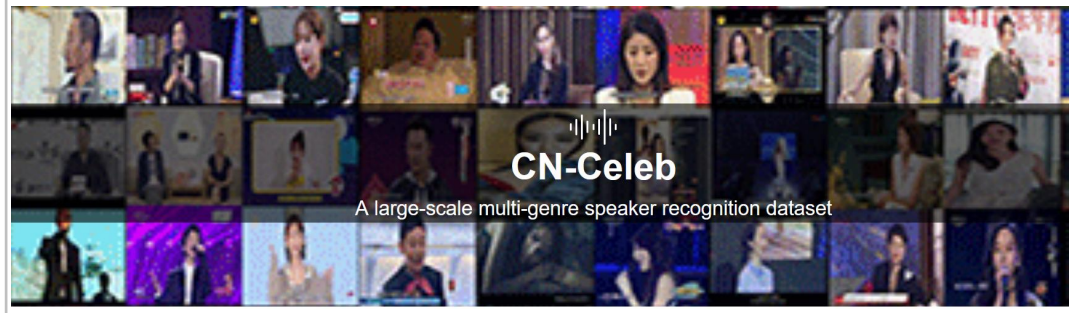
——A Multimodal Corpus of Academic English Speech by Chinese Leaners

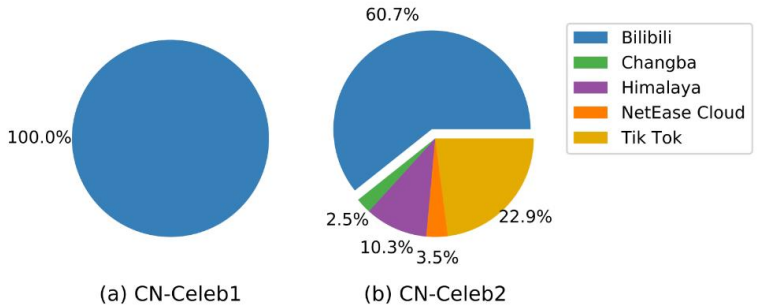
- This corpus consists of **1763** speech samples and it is expanding. Among these students, **1019** were males, and **744** were females.
- **556** students majored in liberal arts, whereas **1179** in STEM majors.
- 3 types of speeches, **informative**, **persuasive**, and **comparative**, were collected.



- This corpus paves the way for further studies on Chinese EFL learners multimodal performance.
- It also contributes to second language teaching and assessing academic English speaking in China.
- It can be the data source for both human rating and the development of automated scoring system.

CNCeleb: Multi-Genre Speaker Recognition Corpus



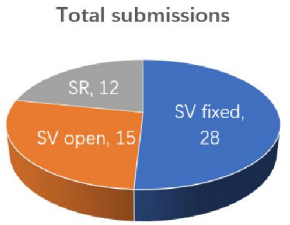


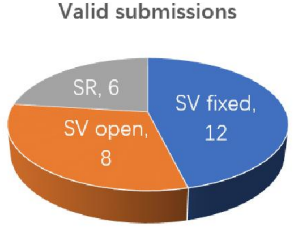
(a) CN-Celeb1 (b) CN-Celeb2

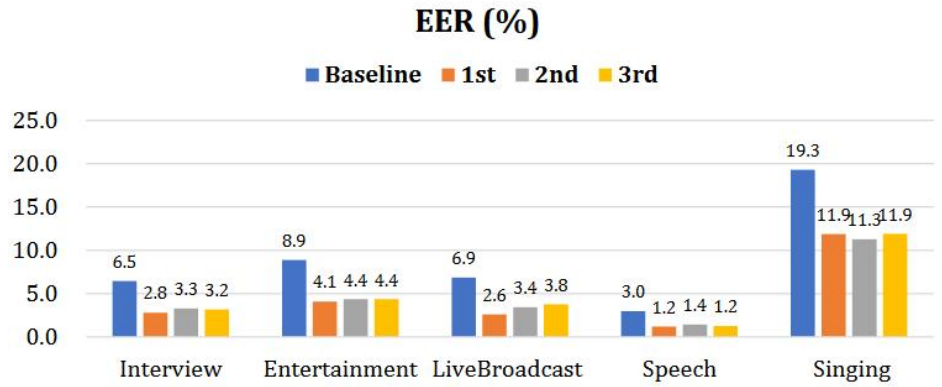
	CN-Celeb1	CN-Celeb2
Language	Chinese	Chinese
Genre	11	11
# of Sources	1	5
# of Spks	1,000	2,000
# of Utters	130,109	529,485
# of Hours	274	1,090
# of SSMC Spks	745	658
Human Check	Yes	Yes

- Large-scale
- Public and free
- Multi-genre
- Cross-genre
- Baseline released

CNCeleb Speaker Recognition Challenge (CNSRC) 2022







希尔贝壳 AISHELL

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.	

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

Open Source	AISHELL-1	400 speakers, 178H	AISHELL-3	218 speakers, 85H
	AISHELL-2	1991 speakers, 1000H	AISHELL-4	60 speakers, 120H

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	

DATATANG www.datatang.com

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