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 26th Conference of the Oriental COCOSDA

4-6 December 2023, Delhi, India

CHINESE DIALECT RESOURCE CONSTRUCTION AT JINAN UNIVERSITY (by prof. Xinzhong LIU)

They focus on the description and research of basic data of Chinese dialects, and currently has three resource platforms that can be queried online.

1

A Comprehensive Dictionary of Chinese Dialects

http://www.fangyanxue.com:8090

This is a dynamic and open Chinese dialectology encyclopedia platform, providing over 3000 dialectology entries, an overview of 54 dialect points, a topic index list of 800 commonly used words, and 702 commonly used word entries, as well as keyword queries.

2

Geographic Information System of Yue, Min and Hakka Dialects

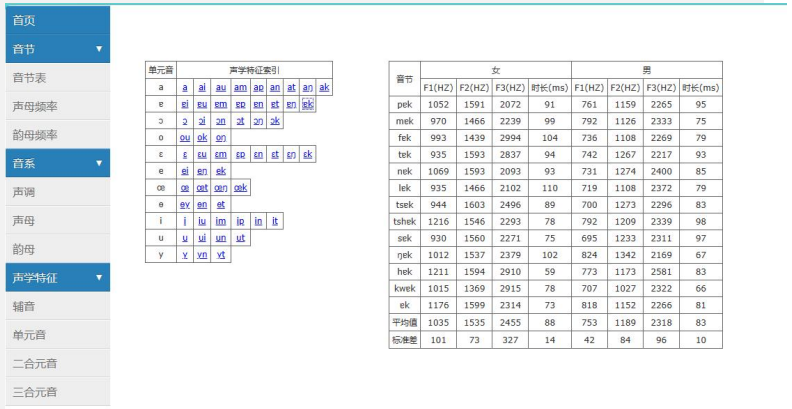
http://1.14.238.88:8099/

This system is a key achievement of the National Social Science Foundation's key project "Construction and Research of Geographic Information System for Yue, Min, and Hakka Dialects". It has uploaded data from 218 points, mainly including single character (divided into ABC three levels), tone sandhi (two character groups), vocabulary, grammar, video, audio, text, and literature. It provides users with functions such as data query, dialect map production, and audio data collection.

3

The Distinctive Features of Chinese Dialects

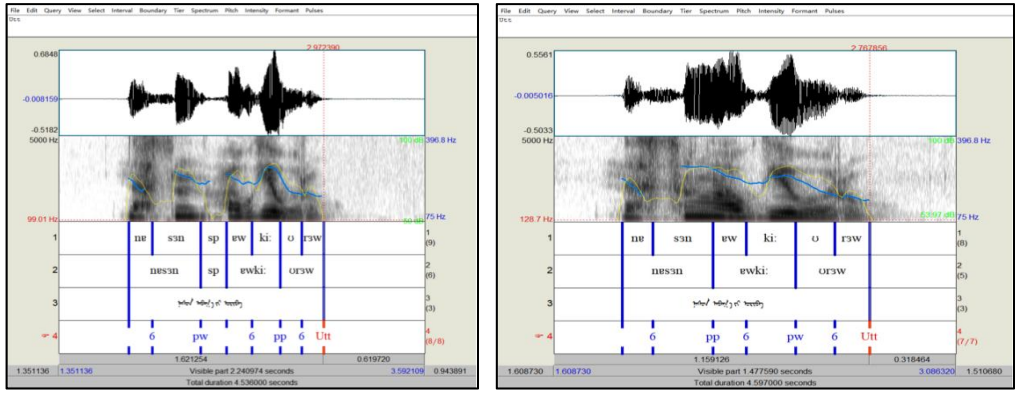
http://www.jnuphon.cn/



This website aims to provide a series of systematic descriptive data on Chinese dialects for researchers, and ultimately completing the construction of a "Chinese dialect phonetics feature database". The first step is to provide basic acoustic data for Cantonese and other dialects, the second step is to provide thematic physiological data, and the third step is to provide some reference data for psycholinguistics and neuroscience in combination with perceptual research.

MONGOLIAN CORPUS FOR INTONATIONAL AND PROSODIC RESEARCH (by prof. AO MIN)

- The corpus consists of **19730** speech samples and still in constructing ...
- Two parts of the database:
  - basic intonatinal patterns (word stress,statement, question, imperative, exclamation) ;
- extanted intonational patterns （discourse intonation,Emotional intonation, Ambiguous intonation,focal accent ）.
- Speakers： 6 native Mongolian speakers, all students, 3 males and 3 Females.
- Annotation:
- Distibution: to be released ....

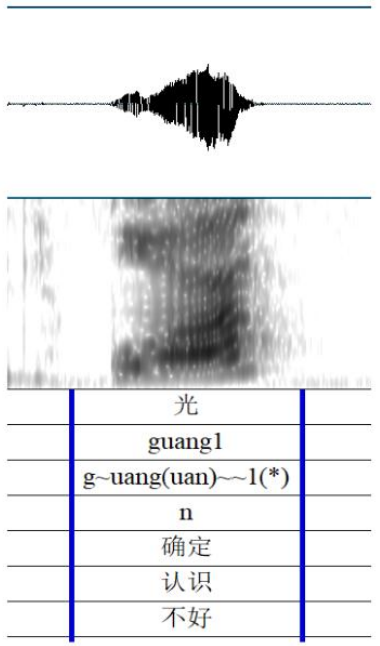


Annotation: ambiguous intonation

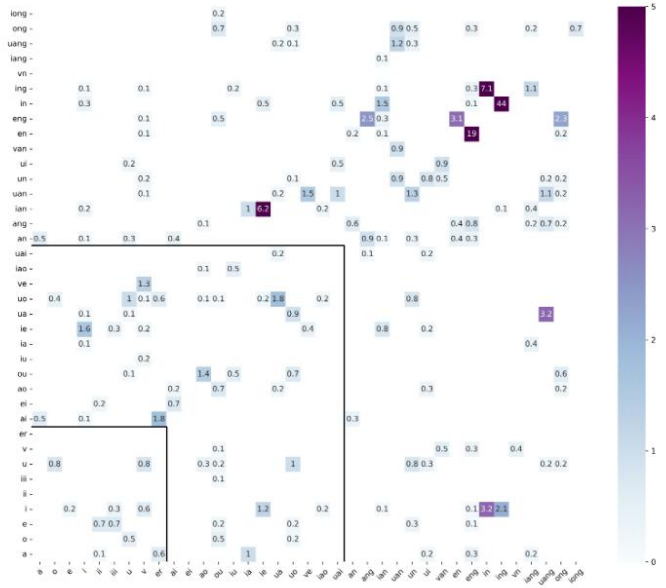
MANDARIN SPEECH DATASET SPOKEN BY UYGHUR CHILDREN

- Speech database consisting of 18,139 Chinese words.
- Recorded through a word naming task for 113 Uyghur children, aged from 4-8 years old.
- All the speech data were phonetically annotated and checked by 3 professional annotaters.
- Collaborating with the Institute of Linguistics, Chinese Academy of Social Sciences
- Open and free, suitable for child language development research

tection.



Annotation



Confusion Matrix

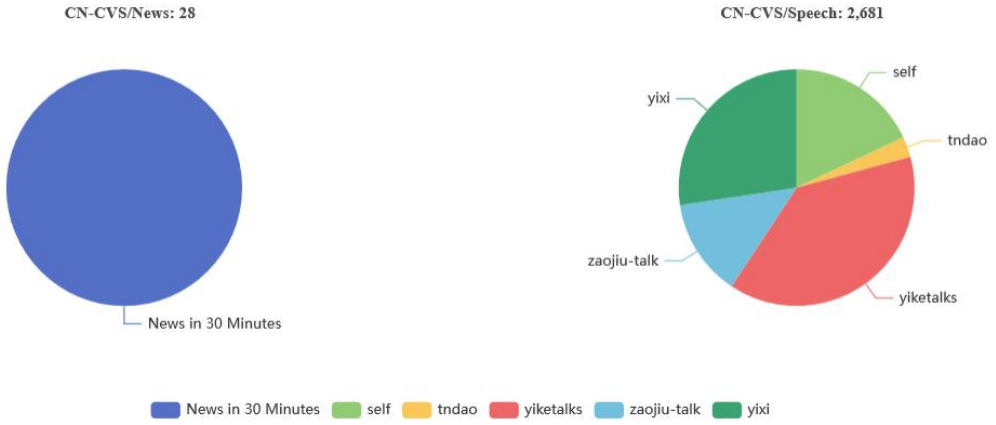
CN-CVS: CHINESE CONTINUOUS VISUAL SPEECH DATASET

2,500 + Speakers

200k + Utterances

300 + Hours

Speakers



CN-CELEB VISUAL SPEECH RECOGNITION CH CHALLENGE (CNVSRC ） 2023

	Fixed Track	Open Track
T1: Single-speaker VSR	CN-CVS, CNVSRC-Single.Dev	No constraint
T2: Multi-speaker VSR	CN-CVS, CNVSRC-Multi.Dev	No constraint

希尔贝壳 AISHELL

海天瑞声

OTHER OPEN SOURCE ACTIVITIES

CNCELAB-AV, Tsinghua University

- CNCeleb Audio-Visual version
- 1,136 speakers, 419k utts, 660+ hours
- Target for audio-visual speaker recognition

XBMU-AMDO31, Northwest Minzu University

- Tibetan Amdo dialect speech
- 31 hours of recordings read by 66 native speakers, 22630 sentences
- Recorded in a quiet room with mobile phones

SHALCAS22A, Shanghai Acoustics Laboratory, CAS and Wuxi Sandu Intelligent Technology Co., Ltd.

- 60 speakers, 14,580 utterances
- Hi-Fi microphone in quiet environment
- Target for numerical password

SLIDESPEECH, Alibaba Group

- 1659 video with synchronized slides from YouTube, 1,000+ hours.
- Covers a variety of domain categories (22 classes)
- Suitable for automatic subtitle generation in online education scen

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