廣東話語音音庫
Cantonese Spoken Language Data Resources

CUCorpora™ is a collection of five different microphone speech corpora. They are the first large scale Cantonese corpora that are publicly available. They also provide important and invaluable infrastructural resources to industrial parties interested in speech enabled products and services.

CUSYL™ and CUWORD™

Descriptions :

CUSYL™ (Version 1.0) - Isolated Cantonese syllables (0:54:59 hour recorded data)

- around 1,800 isolated Cantonese syllables
- cover all valid syllables as well as common lazy and colloquial pronunciations
- 2 male and 2 female speakers (1,800 syllables from each speaker)
- accurate pitch marking provided for 1 male and 1 female

CUWORD™ (Version 1.0) - Cantonese polysyllabic words (32:16:44 hour recorded data)

- around 2,500 polysyllabic Cantonese words
- multiple occurrences for most of the commonly used syllables
- 13 male and 15 female speakers (all 2,500 words from each of them)
- manually verified phonemic transcription provided for every utterance
- designed for training syllable or sub-syllable level acoustic models used in connected Cantonese speech recognition

Recording Condition :

Clean read speech recorded under quiet environment

Deliverables :

CUSYL™	1 CDROM
CUWORD™	5 CDROM

License Fee :

Industrial Research	HK$ 12,000
Academic Research	HK$ 6,000
Commercial Use	HK$ 36,000

CUDIGIT™ and CUCMD™

Descriptions :

CUDIGIT™ (Version 1.0) - Cantonese digit strings (15:06:33 hour recorded data)

- spoken Cantonese digit strings: permutation of **ALL** 1 to 4 digit strings
- supplemented by randomly generated 7, 8 and 16 digit strings
- 50 speakers recorded over several months' time
- manually verified phonemic transcription provided
- designed for building continuous Cantonese digit recognition systems

CUCMD™ (Version 1.0) - Cantonese navigation commands (02:16:19 hour recorded data)

- around 100 Cantonese commands, simulating a navigation control scenario, with alternative and colloquial wordings
- 50 speakers, recorded over several months' time
- ideal for the development of word based command and control products or application systems

Recording Condition :

Clean read speech recorded under quiet environment

Deliverables :

CUDIGIT™	2 CDROM
CUCMD™	1 CDROM

License Fee :

(Industrial Research)	HK$ 5,000
(Academic Research)	HK$ 2,500
(Commercial Use)	HK$ 15,000

廣東話語音音庫
Cantonese Spoken Language Data Resources

CUSENT™

Descriptions :

CUSENT™ (Version 1.0) - Continuous Cantonese sentences (20:26:38 hour recorded data)

- spoken Cantonese continuous sentences corpus designed to be phonetically rich
- 5,100 distinct sentences for training and more than 600 distinct sentences for testing
- 40 male and 40 female speakers participated
- 68 speakers in training set (300 sentences per speaker, a total of 19:24:30 hour recorded data)
- 12 speakers in testing set (100 sentences per speaker, a total of 01:02:08 hour recorded data)
- manually verified phonemic transcription provided
- essential resource for continuous Cantonese speech recognition technology development

Recording Condition :

Clean read speech recorded under quiet environment

Deliverables :

CUSENT™	3 CDROM

License Fee :

(Industrial Research)	HK$ 15,000
(Academic Research)	HK$ 7,500
(Commercial Use)	HK$ 45,000

For the whole package of 12 CDROM in CUCorpora™, there is a discounted price of HK$30,000 for industrial research, or HK$10,000 for non-commercial research and educational purpose in academic institutions, or HK$80,000 for commercial use.

CONTACT

Interested parties are welcome to contact the Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong, for further information.

Attention:	Prof. Tan LEE
Address:	Department of Electronic Engineering, the Chinese University of Hong Kong, Shatin, Hong Kong SAR, China
Fax:	(852) 2603 5558
Email:	tanlee@ee.cuhk.edu.hk
Web page:	http://dsp.ee.cuhk.edu.hk/html/cucorpora.html

This project is conducted under the support from the Industrial Support Fund (AF/20/97). Any opinion, findings, and conclusions or recommendations expressed in this material/event (or by members of the project team) do not reflect the views of the Government of the Hong Kong Special Administrative Region, the Industry Department or the Industry and Technology Development Council.