電話語音音庫
Telephone Speech Corpora (Part I)

CUCall™ is a collection of telephone speech corpora. This collection is the first large-scale Cantonese corpora collected over telephone networks and is made available to the public. There is also a small amount of Mandarin read (words and sentences) and spontaneous speech included. Altogether, these resources provide important and invaluable infrastructural resources to industry as well as the research community. It enables interested parities to deliver speech-enabled products and services over the telephone channels. The reading materials of the corpora are designed with phonetic as well as application-specific considerations.

Cantonese speech data for phonetic coverage

Cantonese Sentences (Version 1.0)
- continuous Cantonese sentences designed to be phonetically rich
- 5,100 different sentences extracted from local newspapers
- over 1,000 speakers
- 30:36:31.14 hour mobile and 30:38:13.42 hour fixed-line network data

Deliverables
- recorded speech data in Microsoft® wave format
- phonemic transcription in Cantonese transcription scheme of the Linguistic Society of Hong Kong
- 2 CDROM for mobile data and 2 CDROM for fixed-line data

License Fee*
- Commercial Use — HK$ 40,000 (mobile) + HK$ 40,000 (fixed-line)
- Research — HK$ 13,500 (mobile) + HK$ 13,500 (fixed-line)

Application specific Cantonese speech data

Cantonese digits (Version 1.0)
- digit strings with lengths range from single digit to 7, 8 and 16 digit long
- cover commonly used digit types such as ID numbers, telephone numbers and credit card numbers
- over 1,000 speakers
- 27:46:54.50 hour mobile and 26:58:58.24 hour fixed-line network data

Cantonese words (Version 1.0)
- listed companies, foreign currencies, navigation commands and names of places in Hong Kong
- useful for enhancing recognition performance in these popular domains
- materials partitioned into small sections and read by hundreds of speakers
- 11:38:46.08 hour mobile and 13:14:16.63 hour fixed-line network data

Deliverables
- recorded speech data in Microsoft® wave format
- phonemic transcription in Cantonese transcription scheme of the Linguistic Society of Hong Kong
- 3 CDROM for mobile data and 3 CDROM for fixed-line data

License Fee*
- Commercial Use — HK$ 24,000 (mobile) + HK$ 24,000 (fixed-line)
- Research — HK$ 7,200 (mobile) + HK$ 7,200 (fixed-line)

For the whole package of 10 CDROM in Part I, there is a discounted price of HK$22,500 for non-commercial research and educational purpose in academic institutions, or HK$32,000 for non-commercial research in non-academic organizations

CUCall™

電話語音音庫
Telephone Speech Corpora (Part II)

Cantonese speech data for different speaking styles

Cantonese Paragraphs (Version 1.0)
- sentences and paragraphs extracted from local newspapers
- paragraphs randomly selected from newspapers
- capturing various speaking behavior in long utterances
- over 1,000 speakers
- 12:54:43.53 hour mobile and 14:37:54.36 hour fixed-line network data

Cantonese Spontaneous Speech (Version 1.0)
- spontaneous speech collected from large number of speakers
- answers to prompted short questions
- intended to capture spontaneous speaking behavior
- over 1,000 speakers
- 8:33:21.42 hour mobile and 9:25:19.43 hour fixed-line network data

Deliverables
- recorded speech data in Microsoft® wave format
- phonemic transcription in Cantonese transcription scheme of the Linguistic Society of Hong Kong
- 2 CDROM for mobile data and 2 CDROM for fixed-line data

License Fee*
- Commercial Use	HK$ 40,000 (mobile) + HK$ 40,000 (fixed-line)
- Research	HK$ 13,500 (mobile) + HK$ 13,500 (fixed-line)

Putonghua speech data

Putonghua Words, Sentences and Spontaneous Speech (Version 1.0)
- words and sentences (including questions) selected from travel and financial domains
- spontaneous answers to prompted short questions
- 6:08:27.84 hour mobile and 5:52:00.14 hour fixed-line network data

Deliverables
- recorded speech data in Microsoft® wave format
- phonemic transcription in pinyin transcription scheme
- 1 CDROM for mobile and fixed-line data

License Fee*
- Commercial Use	HK$ 15,000 (mobile and fixed-line)
- Research	HK$ 5,400 (mobile and fixed-line)

**For the whole package of 5 CDROM in Part II, there is a discounted price of HK$15,000 for non-commercial research and educational purpose in academic institutions, or HK$25,000 for non-commercial research in non-academic organizations*

CUCall™

電話語音音庫
Telephone Speech Corpora

Speaker community

The data were collected from telephone calls of over 1,000 speakers through either the fixed-line or mobile networks. Their ages ranged from 10 to over 60 with a nearly normal distribution centered around twenties.

Recording condition

The speech data is collected over public telephone networks. These comprise of fixed-line and mobile networks. The mobile networks cover GSM, PCS, CDMA as well as TDMA. On the other hand, at the receiving end, the speech signal is captured with common computer telephony card. Speech data are digitized at 8kHz using μ-law companding.

CONTACT

Interested parties are welcome to contact Department of Electronic Engineering[1] and Department of Systems Engineering and Engineering Management[2], the Chinese University of Hong Kong, Hong Kong, for further information.

Attention: Prof. Tan LEE[1] or Prof. Helen MENG[2]
Address: Department of Electronic Engineering[1]/Department of Systems Engineering and Engineering Management[2], the Chinese University of Hong Kong, Shatin, Hong Kong SAR, China
Fax: (+852) 2603 5558[1] / (+852) 2603 5505[2]
Email: tanlee@ee.cuhk.edu.hk or hmmeng@se.cuhk.edu.hk
Web page: http://dsp.ee.cuhk.edu.hk/html/cucall.html

CUCall™