CORPUS INFORMATION FORM

Check (x) or fill in the blank (_____) if appropriate.

1. Name of the Corpus: CUCorpora

2. IPR Holder: Chinese University of Hong Kong

3. Corpus Type:

- ■ Speech Corpus (X)
- ■ Text Corpus ( )

4. If it is a speech corpus:

- Purpose:
  - ■ ASR (X)
  - ■ TTS (X)
  - ■ Other, please specify Bilingual parallel / comparable corpora, e.g. machine translation
- Language:
  - ■ Putonghua (SC) ( )
  - ■ Mandarin in Taiwan (TC) ( )
  - ■ Cantonese in HK (TC) (X)
  - ■ Other, please specify _____
- Style:
  - ■ Read speech (X)
  - ■ Spontaneous speech ( )
  - ■ Conversational speech ( )
  - ■ Other, please specify _____
- Channel:
  - ■ Close-talk Microphone (X)
  - ■ Telephone ( )
  - ■ Mobile phone ( )
  - ■ Other, please specify Desktop microphone
- Sampling Rate: 16000 k Hz
- Sampling Precision:
  - ■ PCM (X), 16 bits per sample
  - ■ A-law ( )
  - ■ Miu-law ( )
  - ■ Other, please specify _____
- Corpus size: ~70 hours ~160 speakers
- SNR level: N.A. dB
- Transcriptions:

- ■ Character tier (SC) ( )
- ■ Character tier (TC) (X)
- ■ Canonical Pinyin tier ( )
- ■ Other canonical pronunciation tier, please specify Cantonese transcription in LSHK format
- ■ Surface form IF tier ( )
- ■ Surface form IPA tier ( )
- ■ Surface form SAMPA-C tier ( )
- ■ Other surface form tier, please specify _____
- ■ Other transcription, please specify monosyllable data from two speakers are manually pitch-marked for pitch-synchronous TTS algorithm and application development
- ■ Other transcription, please specify _____
- ■ Other transcription, please specify _____

5. If it is a text corpus:

- Language:
  - ■ SC ( )
  - ■ TS ( )
  - ■ Other, please specify _____
- Domain:
  - ■ Culture ( )
  - ■ Economy ( )
  - ■ Military ( )
  - ■ News ( )
  - ■ Politics ( )
  - ■ Sciences ( )
  - ■ Sports ( )
  - ■ Other, please specify _____
  - ■ Other, please specify _____
  - ■ Other, please specify _____
  - ■ Other, please specify _____
  - ■ Other, please specify _____
  - ■ Other, please specify _____
- Corpus size: _____ Mega characters
- Tag Information:
  - ■ Word segmentation: ( )
  - ■ Part-of-Speech ( )
  - ■ Other, please specify _____
  - ■ Other, please specify _____
  - ■ Other, please specify _____
  - ■ Other, please specify _____

6. A brief Description of the Corpus:

CUCorpora is the world's first large scale Cantonese spoken language corpora that is

publicly available. CUCorpora is made up of several sub-corpora that are designed for different specific domain of applications. For pitch-synchronous speech synthesis, there are full-set of manually pitch-marked Cantonese syllables (CUSYL). For ASR, there are phonetically-rich speech data (CUWORD and CUSENT) collected from large number of speakers for development of speaker independent ASR. There are also domain-specific data such as digit strings, command and control words that are collected to enhance the performance for continuous digit string recognition as well as command and controlled applications. All of these corpora are 100% manually transcribed at the phonemic level using the transcription scheme defined by the Linguistic Society of Hong Kong (LSHK). Further details, samples or licensing information could be obtained from http://dsp.ee.cuhk.edu.hk/speech/cucorpora.