CORPUS INFORMATION FORM

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

1. Name of the Corpus: CACSC (Cantonese Accent Chinese Speech Corpus)

2. IPR Holder: CST/THU

3. Corpus Type:

- ■ Speech Corpus (x)
- ■ Text Corpus ( )

4. If it is a speech corpus:

- Purpose:
  - ■ ASR (x)
  - ■ TTS ( )
  - ■ Other, please specify _____
- Language:
  - ■ Putonghua (SC) (x)
  - ■ Mandarin in Taiwan (TC) ( )
  - ■ Cantonese in HK (TC) ( )
  - ■ Other, please specify Cantonese Accent
- Style:
  - ■ Read speech (x)
  - ■ Spontaneous speech ( )
  - ■ Conversational speech ( )
  - ■ Other, please specify _____
- Channel:
  - ■ Close-talk Microphone (x)
  - ■ Telephone ( )
  - ■ Mobile phone ( )
  - ■ Other, please specify _____
- Sampling Rate: 16 k Hz
- Sampling Precision:
  - ■ PCM (x), 16 bits per sample
  - ■ A-law ( )
  - ■ Miu-law ( )
  - ■ Other, please specify _____
- Corpus size: ~207 hours 203 speakers 41 CDs ~24GB
- SNR level: _____ dB
- Transcriptions:
  - ■ Character tier (SC) (x)

- ■ Character tier (TC) ( )
- ■ Canonical Pinyin tier (x)
- ■ Other canonical pronunciation tier, please specify _____
- ■ Surface form IF tier ( )
- ■ Surface form IPA tier ( )
- ■ Surface form SAMPA-C tier ( )
- ■ Other surface form tier, please specify _____
- ■ Other transcription, please specify _____
- ■ 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:

CACSC is the first of a series of Chinese speech corpora with different accents. *CACSC* contains 25 Giga Bytes utterances uttered by 104 males and 100 females. Speakers are covered according to the actual proportions of the age (teenager, youngster, middle-aged and elder people) and education (junior high school, senior high school, college,

university) distributions. The design of the prompting text materials are based on the following demands: a) Chinese syllables well balanced, almost all syllables should be covered while more frequently used syllables may appear more often. b) Chinese syllables followed by a suffix that causes a retroflexion of the preceding vowel, typical of the pronunciation of standard Chinese and of some dialects, are considered. c) The reading style is sentence by sentence with punctuation marks read out. The sampling was undertaken at 16KHz rate with 16bit-width data precision through a standard SoundBlaster of a personal computer under ordinary office environment. *CACSC* is mainly based on the standard Chinese, known as Putonghua, with light Cantonese accents. The establishment of *CACSC* offers a testing bed for robust speech recognition of a certain regional accent.