CORPUS INFORMATION FORM

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

1. Name of the Corpus: BIT-MobileSpeech

(Mobile Phone Speech Corpus for Traffic Information Query)

2. IPR Holder:

Modern Communication Lab. Of Beijing Institute Technology

3. Corpus Type:

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

4. If it is a speech corpus:

- Purpose:
  - ■ ASR (x)
  - ■ TTS ( )
  - ■ Other, please specify _____Keyword Spotting_____
- Language:
  - ■ Putonghua (SC) (x)
  - ■ Mandarin in Taiwan (TC) ( )
  - ■ Cantonese in HK (TC) ( )
  - ■ Other, please specify _____
- Style:
  - ■ Read speech (x)
  - ■ Spontaneous speech ( )
  - ■ Conversational speech ( )
  - ■ Other, please specify _____
- Channel:
  - ■ Close-talk Microphone ( )
  - ■ Telephone ( )
  - ■ Mobile phone (x)
  - ■ Other, please specify _____
- Sampling Rate: __8____ k Hz
- Sampling Precision:
  - ■ PCM ( )_____ bits per sample
  - ■ A-law ( )
  - ■ Miu-law ( )
  - ■ Other, please specify _____4-Bit Dialogic ADPCM_____

- Corpus size: ____~2____ hours ____30____ speakers ____~10MB____
- SNR level: ____25~40____ dB
- Transcriptions:
  - ■ Character tier (SC) (x)
  - ■ Character tier (TC) ( )
  - ■ Canonical Pinyin tier ( )
  - ■ 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:

The Mobile Phone Speech Corpus is a read speech corpus designed for Traffic Information Query. There are 30 speakers (15 males and 15 females) by now. Every speaker reads 20 sentences and all sentences are different from each other. Totally, 600 sentences are collected. This corpus has some particular characteristics:

1) The communication network contains not only PSTN but also the GSM or CDMA;
2) The terminal equipments contain many kinds of mobile telephones with different type from different manufacturers.
3) The domain of the corpus is based Traffic Information Query and it is labeled with many keywords of traffic stations, bus lines and locations, etc.

This corpus can be used in many aspects, for example, designing traffic-oriented automatic information service system, keyword spotting, speaker recognition, and so on. And this corpus is still in further extending.