CORPUS INFORMATION FORM

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

1. Name of the Corpus: CSTSC-Flight Corpus (Chinese Spontaneous Telephone Speech Corpus in the flight enquiry and reservation domain)

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 _____
- Style:
  - ■ Read speech ( )
  - ■ Spontaneous speech (x)
  - ■ Conversational speech ( )
  - ■ Other, please specify _____
- Channel:
  - ■ Close-talk Microphone ( )
  - ■ Telephone (x)
  - ■ Mobile phone ( )
  - ■ Other, please specify _____
- Sampling Rate: 8 k Hz
- Sampling Precision:
  - ■ PCM ( ), _____ bits per sample
  - ■ A-law (x)
  - ■ Miu-law ( )
  - ■ Other, please specify _____
- Corpus size: ~100 hours >200 speakers ~3GB

- 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 Spontaneous acoustic phenomena information
  - ■ 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 flight enquiry and reservation domain telephone speech corpus CSTSC-Flight

is taken from the real life. There are about 3 GB raw data. This spontaneous speech corpus can be used to: 1) train a good acoustic model suitable for spontaneous speech recognition because it contains a lot of spontaneous speech phenomena; 2) collect and analyze the spoken sentences (in text) because it contains many spoken language phenomena, which is no doubt useful for natural language parsing; 3) extract the domain-specific knowledge, including domain-specific keywords, key phrases and so on, for domain-specific applications.

*Please note: Only approximately half of the data includes transcriptions.