# CORPUS INFORMATION FORM

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

1. **Name of the Corpus:** **ATR CHRD (**Chinese Hotel Reservation Dialogue**)**

2. **IPR Holder:** **ATR Spoken Language Translation Research Laboratories**

## 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 ( 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: 4 hours 50 speakers
- SNR level: > 30 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:

ATR Spoken Language Translation Research Laboratories has collected a multi-lingual hotel reservation dialogue corpus, which presently consists of Japanese, English and

Chinese speech. The ATR Chinese Hotel Reservation Dialogue (CHRD) database is the Chinese Putonghua parallel speech corpus included in the multi-lingual project. The to-be-contributed part of CHRD includes speech by 50 speakers, 25 males and 25 females. All the speakers were from Beijing area and of 20 ~ 30 years old. Each speaker read four dialogue sides, and a number of common language-parallel sentences. The total number of utterances is about 4,500, and the speech lasts about 4 hours.