CORPUS INFORMATION FORM

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

1. Name of the Corpus: CASS (Chinese Annotated Spontaneous Speech)

2. IPR Holder: CLSP/JHU

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 ( )
  - ■ Mobile phone ( )
  - ■ Other, please specify far field microphone , room acoustics
- Sampling Rate: 16 k Hz
- Sampling Precision:
  - ■ PCM (X), 16 bits per sample
  - ■ A-law ( )
  - ■ Miu-law ( )
  - ■ Other, please specify _____
- Corpus size: 3 hours 7speakers
- 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 (X)
- ■ 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 Chinese Annotated Spontaneous Speech (CASS) corpus contains phonetically transcribed spontaneous speech. This corpus was created to begin to collect samples of most of the phonetic variations in Mandarin spontaneous speech due to pronunciation effects, including allophonic changes, phoneme reduction, phoneme deletion and insertion,

as well as duration changes. It is intended for use in pronunciation modeling for improved automatic speech recognition.