CORPUS INFORMATION FORM

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

1. Name of the Corpus: Wu-Dialectal Chinese Speech Corpus (WDCS)

2. IPR Holder: The Johns Hopkins University

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) ( )
  - ■ Mandarin in Taiwan (TC) ( )
  - ■ Cantonese in HK (TC) ( )
  - ■ Other, please specify Wu-Dialectal Chinese Speech, or Speech Corpus of Chinese influenced by native Wu Dialect
- Style:
  - ■ Read speech (x)
  - ■ Spontaneous speech (x)
  - ■ 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: 11 hours 100 speakers

- 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 (x)
  - Surface form IPA tier ( )
  - Surface form SAMPA-C tier ( )
  - Other surface form tier, please specify Sound Changes Representation using SAMPA-C
  - Other transcription, please specify Non-speech information
  - 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 WDC speech corpus contains 11 hours speech data, including 5.5 hours' read speech data and 5.5

hours' spontaneous speech data, uttered by 100 native Shanghai speakers. The prompting texts were designed by using an automatic sentence selection algorithm so as to cover the Chinese language phenomena phonetically as much as possible for read speech. And 5 topics, including Sports, Politics & Economy, Entertainment, Lifestyle and Technologies, as well as some corresponding sub-topics are elaborately designed for speakers to talk when recording the spontaneous speech. The speech data were recorded by using high quality headset microphones, and stored in 16k 16bit standard windows PCM format. With four-level manual transcriptions, such as Chinese character layer, canonical Chinese pinyin layer, surface form Chinese IF layer and some miscellaneous layer, this corpus could be a good one for Chinese language processing.