CORPUS INFORMATION FORM

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

1. Name of the Corpus: 500-People TRSC (Telephone Read Speech Corpus)

2. IPR Holder: d-Ear

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 ( )
  - ■ Telephone (x)
  - ■ Mobile phone ( )
  - ■ Other, please specify _____
- Sampling Rate: 8 k Hz
- Sampling Precision:
  - ■ PCM ( ), _____ bits per sample
  - ■ A-law ( )
  - ■ Miu-law (x)
  - ■ Other, please specify _____
- Corpus size: 72 hours 500 speakers ~2GB
- 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 _____
  - ■ 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:

This is a large telephone read speech corpus by about 500 speakers, half male and half female. Every speaker reads about 110 sentences most of which are taken from *the China*

People's Daily newspapers. It also contains some sentences to cover the connected digits and English characters. This corpus has some characteristics:

1) Roughly di-IF (di-Initial/Final) balanced;
2) Digital and English character balanced;
3) Large number of speakers;

This corpus can be used to train the acoustic model on telephone channel for LVCSR. It can also be used to do many other research works, for example, speaker identification and verification, and so on.