

## CORPUS INFORMATION FORM

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

1. Name of the Corpus: BIT-TeleSpeech

(telephone read speech corpus)

2. IPR Holder:

Modern Communication Lab. of Beijing Institute Technology

3. Corpus Type:

- Speech Corpus (x )
- Text Corpus ( )

4. If it is a speech corpus:

- Purpose:

- ASR ( )
- TTS ( )
- Other, please specify speaker recognition

- 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 (x ), 16 bits per sample
- A-law ( )
- Miu-law ( )
- Other, please specify \_\_\_\_\_

- 
- Corpus size: 0.5 hours 15 speakers 30MB
  - SNR level: ~35 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 \_\_\_\_\_

### 5. If it is a text corpus:

- Language:
  - SC ( )
  - TS ( )
  - Other, please specify \_\_\_\_\_
- Domain:
  - Culture ( )
  - Economy ( )
  - Military ( )
  - News ( )
  - Politics ( )
  - Sciences ( )
  - Sports ( )
  - Other, please specify \_\_\_\_\_
- Corpus size: \_\_\_\_\_ Mega characters
- Tag Information:
  - Word segmentation: ( )
  - Part-of-Speech ( )
  - Other, please specify \_\_\_\_\_

### 6. A brief Description of the Corpus:

This is a telephone read speech corpus by 15 speakers (8 males and 7 females). Every speaker reads the same 40 sentences and each of the sentences is about 3s.

This corpus can be used for speaker recognition.