CORPUS INFORMATION FORM

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

1. Name of the Corpus: TNDC (Telephone Name Dialing 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 Keyword Spotting
- 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: 60 hours 400 speakers ~1.7GB
- 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:

The Telephone Name Dialing Corpus is a read speech corpus designed for name-dialing system. There are about 500 Chinese names and 50 sentence templates chosen for designing the target sentences. Every speaker reads about 110 sentences, including 70 sentences and 40 isolated names. Totally, 400 speaker's speech data are collected. This

corpus can be used in many aspects, for example, name dialing system, keyword spotting, speaker recognition, and so on.