CORPUS INFORMATION FORM

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

1. Name of the Corpus: BIT-TonalName

(Tonally Confusing Name 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 (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 (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: 5.4 hours 20 speakers 607MB
- SNR level: 30~50 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 corpus is collected for the study of Chinese name recognition and the influence of tonal information.

The vocabulary is composed of 50 pairs of short names (bi-syllabic names) and 50 pairs of long names (tri-syllabic names). Within a pair the two names are tonally confusing, which means that they have the same base syllables (syllable without tone) but different tones.

When designing the vocabulary the following factors were taken into account:

1. The corpus covers all 21 initials and 38 finals of Mandarin.
2. Tonally confusing syllables are assigned to different position in these name pairs: at the beginning, in the middle or at the end.
3. Among the most popular 45 surnames in Chinese names, the corpus covers 39 ones.

The corpus was contributed by 10 male and 10 female speakers, and all the speakers are students with native language of Mandarin. Each speaker was required to utter the 200 names twice in a quiet environment. So there are 20*200*2 = 8,000 utterances in this corpus.