CORPUS INFORMATION FORM

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

1. Name of the Corpus: CCC-VPR3C2005 (CCC 3-Channel Corpus for Text-Ind/-Dep Voiceprint Recognition 2005)

2. IPR Holder: Chinese Corpus Consortium (CCC)

3. Corpus Type:

- ■ Speech Corpus (x)
- ■ Text Corpus ( )

4. If it is a speech corpus:

- Purpose:
  - ■ ASR ( )
  - ■ TTS ( )
  - ■ Other, please specify Voiceprint Recognition (VPR)
- 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: 48 k Hz
- Sampling Precision:
  - ■ PCM (x), 16 bits per sample
  - ■ A-law ( )
  - ■ Miu-law ( )
  - ■ Other, please specify _____
- Corpus size: 69 hours 100 speakers
- SNR level: 30 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 was designed for voiceprint recognition (VPR) or speaker recognition tasks. It contains two subsets, one for text-independent (TI) VPR and the other for text-dependent (TD) VPR.

This corpus can also be used for multi-channel or cross-channel VPR research, because each sentence (in Chinese) was recorded through three different types of microphones simultaneously. The three types of microphones are labeled with 'U', 'L', and 'R', respectively.

All samples are stored in Microsoft wave format files with 48 kHz sampling rate, 16 bit PCM, and mono-channel.

TI-VPR Subset

This dataset contains speech samples uttered by 100 speakers, each of which uttered the same 60 Chinese sentences. The Chinese syllable layer transcription of these sentences is given in file 'text-independent-transcript.txt'. Details of the 100 speakers are listed in the following table.

Speaker ID	***Gender***	***Age***
001-015	male	10~19
016-030	male	20~29
031-045	male	30~39
046-050	male	40~49
051-065	female	10~19
066-080	female	20~29
081-095	female	30~39
096-100	female	40~49

Sample data are stored under the folder 'Data\Text-Independent', and the file names have the following format,

[*speaker id*]_[*sentence number*]_[*microphone type letter*].wav

For instance, file '006_28_U.wav' stands for a wave file of sentence No. 28 uttered by speaker No. 006 recorded through microphone 'U'.

TD-VPR Subset

This dataset contains speech samples uttered by the same 100 speakers as in the TI-VPR Subset. Each speaker uttered 20 short sentences or phrases for 4 times, where, for example, the first 3 times' utterances can be used for training while the remaining for testing. In addition, each speaker uttered another 24 short sentences or phrases for 1 time, which can be used as impostor samples.

Each of the short sentences or phrases contains 5 to 11 syllables. The transcription is given in file 'text-dependent-transcript.txt'.

Samples are stored under the folder 'Data\Text-Dependent', and the file names have the following format,

[*speaker id*]_[*sentence number*]_[*utterance number*]_[*microphone type letter*].wav

For instance, file '010_24_3_R.wav' stands for a wave file of sentence No. 24 uttered by speaker No. 010 for the 3rd time through microphone 'R'.