

CORPUS INFORMATION FORM

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

1. Name of the Corpus: CCC-VPR2C2006-10000 (CCC 2-Channel
Corpus for Voiceprint Recognition 2006 - 10000 speakers)

2. IPR Holder: _____

3. Corpus Type:

- Speech Corpus (x)
- Text Corpus ()

4. If it is a speech corpus:

- Purpose:
 - ASR (x)
 - TTS ()
 - Other, please specify Channel Robust Voiceprint 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 (x)
 - 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: 167 hours 10000 speakers
- SNR level: >30 dB

- Transcriptions:
- Character tier (SC) ()
 - 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 _____
 - 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 contains speech from 1,000 male speakers aged 18-23, each of whom was required to utter 1) a brief introduction (personal information) and 2) 40 Chinese sentences (as listed in the

following table) in the given order. All utterances were required to be made twice, speaking clearly and naturally without any attempt to disguise the voice. For each speaker, the first time the utterance was made it was recorded through a GSM mobile phone and the second time it was recorded through a landline telephone. Only the first 70 seconds of each utterance were recorded.

姓名：××× 年龄：×× 学号：XXX

- | | |
|---------------|--------------|
| 1、疾病给人带来了痛苦 | 21、姐姐帮助弟弟做功课 |
| 2、星期一我得了感冒 | 22、我热爱警察这个职业 |
| 3、八月的天气很闷热 | 23、文明人每天都刷牙 |
| 4、性急容易发生错误 | 24、眼泪有时会流进鼻子 |
| 5、阿姨的女儿在照镜子 | 25、电灯比油灯进步多了 |
| 6、一般的雨衣都漏水 | 26、爱情绝不是娱乐 |
| 7、女工生产黑色的书包 | 27、西北的橘子是南方的 |
| 8、书包里是图书杂志 | 28、这是全自动洗衣机 |
| 9、孩子们到野外去放风筝 | 29、司令员说话真干脆 |
| 10、现在点油灯的少多了 | 30、文艺要为工农兵服务 |
| 11、黄瓜是一种果类蔬菜 | 31、他赛跑得了第一 |
| 12、现在过年也不磕头了 | 32、我给话剧演员化妆 |
| 13、佛教是一种精神信仰 | 33、先生答复学生的问题 |
| 14、口渴了就要喝水 | 34、编辑不做校对工作 |
| 15、他的表现太软弱了 | 35、姐姐把功课做完了 |
| 16、那个医生说姓罗 | 36、公共汽车上也有电灯 |
| 17、医生的工作是治疗疾病 | 37、讲地理最好有模型 |
| 18、今年果树长得很茂盛 | 38、黄瓜不是黄色的 |
| 19、大家都愿意用自来水 | 39、每人都有一把椅子 |
| 20、他的行为应该受到批评 | 40、新刊物在元旦出版 |