A Country Report - COCOSDA Activities in China

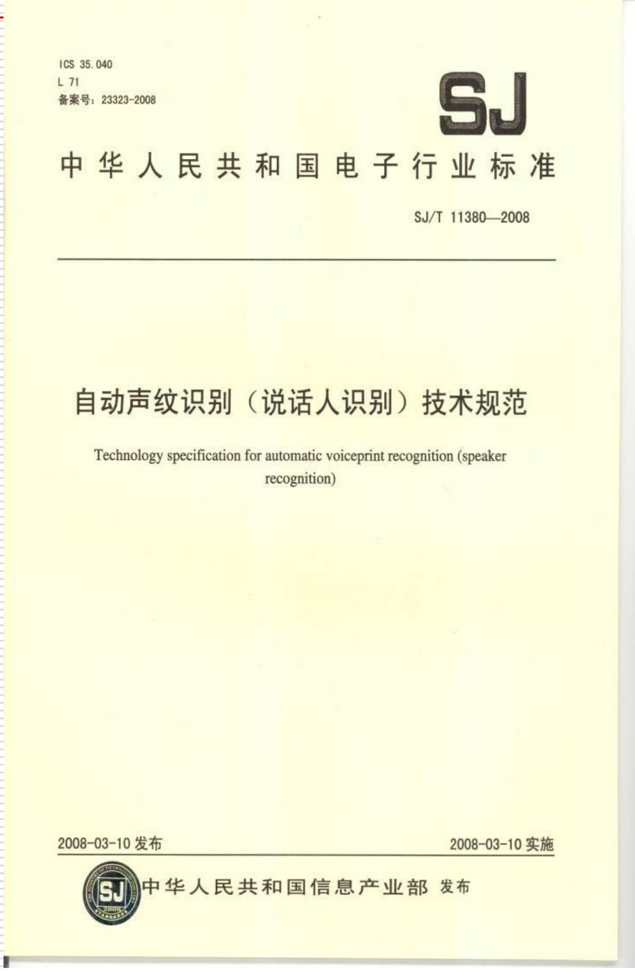
Aijun Li and Thomas Fang Zheng

O-COCOSDA'2008
Kyoto, Japan

25 Nov 2008

Activities - CCC

- **Technology specification for automatic voiceprint recognition (speaker recognition)** (SJ/T 11380-2008) was issued by former Ministry of Information Industries (MII) on Mar 10, 2008.
  - The Standard includes three parts: Terminologies and definitions, data exchange format, and APIs
  - Drafters: CSLT Tsinghua University (CCC co-founder), d-Ear Technologies (CCC co-founder), and China Electronic Standardization Institute (CESI)
  - http://www.CCCForum.org



ICS 35.040
L 71
备案号：23323-2008

SJ

中华人民共和国电子行业标准

SJ/T 11380—2008

自动声纹识别（说话人识别）技术规范

Technology specification for automatic voiceprint recognition (speaker recognition)

2008-03-10 发布　　2008-03-10 实施

中华人民共和国信息产业部 发布

Activities - ChineseLDC

- Working Group meeting was held in the Institute of Linguistics, CASS, June 13th, 2008.

- Corpus sharing situation: 44 corpora were distributed in the last year, covering 15 affiliations (2 domestic companies and 6 international companies)

New Resources

- **CCC-VPR2C2006-10000**
  - 10,000 speaker telephone and mobile phone channels corpus for Voiceprint Recognition finished finally

- **Song’s Melody Corpus**
  - 5,000 song for research on the Query by Humming (QBH) system
  - Transcriptions: musical notation, pitch, duration, tag of sentence start in xml format

- CLDC-LAC-2008-001
  - Chinese Web 5-gram Corpus for statistical language modeling
- CLDC-LAC-2008-002
  - Words and Phrases Corpus for Official Affairs with Occurrence Frequency. (By ZHU Dingfu)
- 2007-863-001
  - Evaluation corpus for SSMT2007, including Chinese-English and English-Chinese Parallel Data, guidelines and reports, and software (by Institute of Computation, CAS)
- CLDC-SPC-2007-001 RASC863- Part II
  - 6 Regional (Changsha, Luoyang, Nanchang, Nanjing, Taiyuan, Wenzhou) accented Mandarin corpus, with Pinyin and orthographical annotation, spontaneous/read speech and selected dialectical words.
- CLDC-LAC-2006-001-004:
  - modern Chinese balanced raw text corpora from Ministry of Education, tree bank corpus with Lexical for Parsing

- □ Institute of Linguistics, CASS
  - Tibetan speech corpus for synthesis(863 corpus)
  - EMA speech and articulatory corpus with 4 Mandarin speakers.
  - EPG speech corpus for normal speakers and patients speakers after larynx surgery

- □ Language Institute of the Central University for Nationalities: Speech Corpus of Endangered Chinese Minority Languages
  - The speech materials of these two sub-corpus consist of five parts: phonetic sounds, vocabulary, sentence, conversation, and long speech material. Folklores, poems and folk songs, 12 sections of Tujia language and 6 sections of Gelao language are collected.