Shuichi Itahashi and Chiu-yu Tseng Eds. "Computer Processing of Asian Spoken Languages," 1st Edition, Consideration Books, Los Angeles, USA, March 2010, pp. 65-68

THREE
(2) The Chinese Corpus Consortium (CCC)

Thomas Fang Zheng

1 Introduction

In this section we introduce the Chinese Corpus Consortium (CCC, http://www.CCCForum.org), a speech and linguistics resources consortium, as well as currently available resources and CCC-sponsored activities.

The CCC was established dedicated to the Chinese language, and is similar to several existing consortiums and associations such as the Linguistic Data Consortium (LDC)[1], the European Language Resources Association (ELRA)[2], and the Gengo-Shigen-Kyokai Language Resource Association (GSK)[3].

In March 2004, CCC was co-founded by the following universities, institutes, and companies:

a. Beijing d-Ear Technologies Co., Ltd. (d-Ear)
b. The Center for Speech Technology, Tsinghua University (CST*)
c. Institute of Linguistics, Chinese Academy of Social Sciences (CASS)
d. Human Computer Interaction and Multimedia Lab, Tsinghua University (HCI&MM)
e. The Chinese & Oriental Language Information Processing Society in Singapore (COLIPS)
f. ATR Spoken Language Translation Research Labs (ATR)
g. The Center for Language & Speech Processing, The Johns Hopkins University (JHU)
h. The Chinese University of Hong Kong (CUHK)

The aim of the CCC is to provide corpora for Chinese automatic speech recognition, text-to-speech, natural language processing, perception analysis, phonetics analysis, linguistic analysis, and other related tasks. The corpora can be speech- or text-based; read or spontaneous; wideband or narrowband; standard or dialectal Chinese; clean or with noise; or any other kind consistent with the aims of the Consortium.

The purposes of the CCC include the following:

- The collection and integration of existing Chinese speech and linguistic corpus resources, and the continuing creation of new such resources
- The integration and enhancement of existing tools for the creation, transcription, and analysis of Chinese speech and linguistic corpus resources, and the creation of new such tools.

* It has been renamed as the Center for Speech and Language Technologies (CSLT, http://cslt.riit tsinghua.edu.cn) since 2007.

- The collection, organization, and introduction of specifications and standards for Chinese speech and language research and development
- The promotion of the exchange of Chinese speech and linguistic corpus resources

Headquartered in Beijing, China, the CCC is supported by the Chinese Language Resources branch of the High-tech Enterprises Association of the Beijing Experimental Zone for the Development of New Technology Industries (HTEA)[4] and is administered by d-Ear Technologies, which works for the mutual promotion of the standardization and industrialization of Chinese language resources.

Currently, the CCC board includes Council Chair Dr. Thomas Fang Zheng (Center for Speech and Language Technologies, Tsinghua University), Vice Council Chair Dr. Satoshi Nakamura (Advanced Telecommunications Research), and Standing Secretary Dr. Yi Liu (Center for Speech and Language Technologies, Tsinghua University).

2 Available Corpora and Resources

As listed in Table 1, the CCC currently has over 30 corpora for various research purposes. In addition, the CCC also maintains the free resources (for both members and non-members) and member resources (available to all current members) listed in Table 2.

TABLE 1 CCC CORPORA

Corpus Name	Owner	Description
CHRD	ATR	Chinese Hotel Reservation Dialogue
SCSC	CASS	Syllable Corpus of Standard Chinese
WCSC	CASS	Word Corpus of Standard Chinese
ASCCD	CASS	Annotated Speech Corpus of Chinese Discourse
CADCC	CASS	Chinese Annotated Dialogue and Conversation Corpus
PSC LT	COLIPS	Singapore Primary School Chinese Language Text
CACSC	CST	Cantonese Accent Chinese Speech Corpus
CSTSC-Flight Corpus	CST	Chinese Spontaneous Telephone Speech Corpus (flight enquiry/reservation domain)
CUCorpora	CUHK	Cantonese spoken language corpora
TRSC	d-Ear	500-Person Telephone Read Speech Corpus
TNDC	d-Ear	Telephone Name Dialing Corpus
CoSS-0	HCI&MM	TH Corpus of Speech Synthesis No.0
CASS Corpus	JHU	Chinese Annotated Spontaneous Speech
WDCS Corpus	JHU	Wu-Dialectal Chinese Speech
BIT-MobileSpeech	BIT	Mobile Phone Speech Corpus for Traffic Information Query
BIT-Mobile Talk	BIT	Mobile Phone Conversational Speech Corpus for Travel
BIT-TeleSpeech	BIT	Telephone Read Speech Corpus
BIT-TonalName	BIT	Tonally Confusing Name Speech Corpus
BIT-MonoSyllable	BIT	Mandarin Mono-Syllable Corpus
CCC-VPR27C2006-50	CCC	CCC 27-Channel Corpus for Voiceprint Recognition -50 speakers
CCC-VPR36C2006-100	CCC	CCC 36-Channel Corpus for Voiceprint Recognition - 100 speakers
CCC_AC2006_ASR	CCC	Affective Speech Recognition database
CCC_AC2006_CAR	CCC	Chinese Affect Recognition database
CCC_AC2006_FEE	CCC	Face Emotional Expression database
CCC_AC2006_NPF	CCC	Nice Pose Faces Database
CCC_AC2006_MMA	CCC	Multi-Modality Affect Database
CCC_AC2006_MMA_P	CCC	Partial Multi-Modality Affect Database

TABLE 1 CCC CORPORA [CONTINUED]

Corpus Name	Owner	Description
CCC-VPR3C2005	CCC	CCC 3-Channel Corpus for Voiceprint Recognition
CCC-VPR2C2005-1000X	CCC	CCC 2-Channel Corpus for Voiceprint Recognition - 11 kHz
CCC-VPR2C2005-1000	CCC	CCC 2-Channel Corpus for Voiceprint Recognition - 1000 speakers
CCC-VPR2C2005-3000	CCC	CCC 2-Channel Corpus for Voiceprint Recognition - 3000 speakers
CCC-VPR2C2005-6000	CCC	CCC 2-Channel Corpus for Voiceprint Recognition - 6000 speakers
CCC-VPR2C2006-10000	CCC	CCC 2-Channel Corpus for Voiceprint Recognition - 10000 speakers
CCC-VPR27C2006-50	CCC	CCC 27-Channel Corpus for Voiceprint Recognition - 50 speakers
CCC-VPR36C2006-100	CCC	CCC 36-Channel Corpus for Voiceprint Recognition - 100 speakers
CCC_AC2006_ASR	CCC	Affective Speech Recognition database
CCC_AC2006_CAR	CCC	Chinese Affect Recognition database
CCC_AC2006_FEE	CCC	Face Emotional Expression database
CCC_AC2006_NPF	CCC	Nice Pose Faces Database
CCC_AC2006_MMA	CCC	Multi-Modality Affect Database
CCC_AC2006_MMA_P	CCC	Partial Multi-Modality Affect Database

TABLE 2 FREE AND MEMBER RESOURCES

Resource	Type	Description
Pinyin Syllable List	Free	The complete list of 416 Chinese syllables
Pinyin XIF List	Free	An extended list of pinyin initials and finals
Sampa-C Reference	Free	Reference for the Chinese segmental labeling convention Sampa-C
O-COCOSDA 98-02	Free	Oriental COCOSDA conference proceedings for 1998 through 2002
Word List	Member	An automatically generated wordlist of 50,000 Chinese words with pinyin and count information
CCC-VPR3C2005	Member	CCC 3-Channel Corpus for Voiceprint Recognition
CCC-VPR2C2005-1000	Member	CCC 2-Channel Corpus for Voiceprint Recognition - 1000 speakers

3 Activities Sponsored by CCC

Four of the CCC co-founders—d-Ear, CST, CASS, and HCI&MM—are also co-founders of the Chinese Speech Interactive Technology Group (CSITSG)[5], approved by the Science and Technology Department of the Ministry of Information Industry (MII) of China. In December 2006, the 3rd Symposium of CSITSG and Voiceprint Recognition Standard Approval Meeting was held at Tsinghua University. Dr. Thomas Fang Zheng, the leader of Voiceprint Recognition Special Topic Group, gave a report titled "Statement on the Constituting of Automatic Voiceprint Recognition (Speaker Recognition) Technology Standard." After a full and deep discussion, this standard was approved; the text of the standard has been approved and was announced by MII in March 2008. This was the first standard on voiceprint recognition in China.

CCC also chaired a special session on speaker recognition during the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP'2006) held in Singapore during December 12–17, 2006. The purpose was to invite researchers in this field to present their state-of-the-art technical achievements, and to provide a platform for developers in this field to evaluate their speaker recognition systems using the same CCC-provided database. Eight sites from seven countries/regions participated in the CCC SRE 2006.

References

1. LDC, Linguistic Data Consortium, Philadelphia, USA. http://www.ldc.upenn.edu/.
2. ELRA, European Language Resources Association, Paris, France. http://www.elra.info/.
3. GSK, Gengo-Shigen-Kyokai Language Resource Consortium, Tokyo, Japan. http://www.gsk.or.jp/.
4. HTEA, High-tech Enterprises Association of the Beijing Experimental Zone for the Development of New Technology Industries, Beijing, China, http://www.htea.net.cn/ .
5. CSITSG, Chinese Speech Interactive Technology Standard Group, Anhui, China. http://www.speechstandard.org.cn/.
6. Zheng, T. F., Song, Z.-J., Zhang, L.-H., Brasser, M., Wu, W. and Deng, J. CCC speaker recognition evaluation 2006: overview, methods, data, results and perspective. in Huo, O., Ma, B., Chng, E.-S. and Li, H. (eds.) Chinese Spoken Language Processing (Springer-Verlag Berlin Heidelberg, Germany) pp. 485-493. 2006