CCC Speaker Recognition Evaluation 2006: Overview, Methods, Data, Results and Perspective

Thomas Fang Zheng[1,2], Zhanjiang Song[2,3], Lihong Zhang[3], Michael Brasser[1,3], Wei Wu[2], and Jing Deng[2]

[1] Chinese Corpus Consortium (CCC)
fzheng@tsinghua.edu.cn, mbrasser@d-Ear.com
http://www.CCCForum.org
[2] Center for Speech Technology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, 100084
{fzheng, szj, wuwei, dengj}@cst.cs.tsinghue.edu.cn
[3] Beijing d-Ear Technologies Co., Ltd.
{zjsong, lhzhang, mbrasser}@d-Ear.com
http://www.d-Ear.com

Abstract. For the special session on speaker recognition of the *5th International Symposium on Chinese Spoken Language Processing* (ISCSLP 2006), the *Chinese Corpus Consortium* (CCC), the session organizer, developed a speaker recognition evaluation (SRE) to act as a platform for developers in this field to evaluate their speaker recognition systems using two databases provided by the CCC. In this paper, the objective of the evaluation, and the methods and the data used are described. The results of the evaluation are also presented.

Keywords: Speaker recognition, Evaluation.

1 Introduction

Speaker recognition (or voiceprint recognition, VPR) is an important branch of speech processing with applications in many fields, including public security, anti-terrorism, forensics, telephony banking, and personal services. However, there are still many fundamental and theoretical problems to solve, such as issues with background noise, cross-channel recognition, multi-speaker recognition, and difficulties arising from short speech segments for training and testing [1-3].

In addition to inviting researchers to present their state-of-the-art achievements in various aspects of the speaker recognition field, this special session on speaker recognition of the *5th International Symposium on Chinese Spoken Language Processing* (ISCSLP 2006) provides a platform for VPR developers to evaluate their speaker recognition systems using two databases provided by the *Chinese Corpus Consortium* (CCC). This paper is organized as follows. In Section 2, an overview of

Q. Huo et al. (Eds.): ISCSLP 2006, LNAI 4274, pp. 485 – 493, 2006.

the evaluation is given. Details of the evaluation are described in Section 3. The summary and further perspectives on the evaluation are given in Section 4.

2 Overview of the Evaluation

2.1 Organizer

This speaker recognition evaluation (SRE) was organized by the CCC. The CCC was founded in March 2004, sponsored by Dr. Thomas Fang Zheng and co-founded by 8 universities, institutes and companies. The aim of the CCC is to provide corpora for Chinese ASR, TTS, NLP, 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 of any other kinds which are deemed helpful for the aforementioned purposes. Currently there are numerous corpora available from the CCC. For more information, readers can refer to the official website of the CCC (http://www.CCCForum.org) and paper [4].

2.2 Objective

The purpose of this SRE is to provide an opportunity for VPR researchers and developers to exchange their ideas and to help push forward, especially, corresponding work on Chinese language data. It can be seen as a specially focused event, similar to other well-known events (e.g. the speaker recognition evaluations carried out by NIST [5-7]).

3 The CCC 2006 SRE

Detailed information on the CCC 2006 SRE is given in this section.

3.1 Task Definition

The CCC 2006 SRE covers the following six tasks:

1) Text-dependent single-channel speaker verification.
2) Text-independent single-channel speaker verification.
3) Text-dependent cross-channel speaker verification.
4) Text-independent cross-channel speaker verification.
5) Text-independent single-channel speaker identification.
6) Text-independent cross-channel speaker identification.

All of the above tasks are optional for participants.

Please note that for text-dependent speaker-verification tasks in this evaluation (both single-channel and cross-channel), a test sample is treated as a true speaker trial only when both the speaker identity and the content match those of the training samples.

3.2 Performance Measure

The methods for measuring the performance of the participating systems are described below.
(1) Speaker Verification

The performance of a speaker verification system is evaluated by a Detection Error Tradeoff (DET) curve and a detection cost function (C_{Det}) [6]. The $C_{Det}$is defined as a weighted sum of miss and false alarm error probabilities:

$$C_{Det} = C_{Miss} \times P_{Miss} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm} \times \left(1 - P_{Target}\right) \tag{1}$$

where C_{Miss} and $C_{FalseAlarm}$ are the relative costs of miss errors and false alarm errors, and P_{Target} is the *a priori* probability of the specified target speaker (in this evaluation, these parameters are set as in Table 1). P_{Miss} and $P_{FalseAlarm}$ are the miss probability and false-alarm probability, respectively. A miss error occurs when a true speaker model of a test segment is rejected, while a false alarm error occurs when an impostor model of a test segment is accepted. The miss probability is defined as

$$P_{Miss} = \frac{N_{Miss}}{N_{VS}} \times 100\% \tag{2}$$

where N_{Miss} is the number of miss errors and N_{VS} is the number of true **s**peaker trials. The false alarm probability is defined as

$$P_{FalseAlarm} = \frac{N_{FalseAlarm}}{N_{VI}} \times 100\% \tag{3}$$

where $N_{FalseAlarm}$ is the number of false alarm errors and N_{VI} is the number of impostor trials.

Table 1. Speaker verification cost model parameters

C_{Miss}	$C_{FalseAlarm}$	P_{Target}
10	1	0.05

(2) Speaker Identification

The performance of a speaker identification system is evaluated by its *Identification Correctness Rate* (P_{IC}), which is defined as:

$$P_{IC} = \frac{N_{IC}}{N_{IT}} \times 100\% \tag{4}$$

where N_{IC} is the number of correctly identified segments. A correctly identified segment means that the system should output the model speaker's identity as

top-candidate for "in-set" tests, and output a "non-match" flag for "out-of-set" tests. N_{IT} is the total number of trial segments.

3.3 Corpora

The data sets, including development data and evaluation data, were extracted from two CCC databases, CCC-VPR3C2005 and CCC-VPR2C2005-1000.

CCC-VPR3C2005: This corpus contains two subsets, one for text-independent VPR and the other for text-dependent VPR. This corpus can also be used for multi-channel or cross-channel VPR research, because each sentence (in Chinese) was simultaneously recorded through three different types of microphones. The three types of microphones are labeled with 'U', 'L', and 'R', respectively. All samples are stored in Microsoft wave format files with a 48 kHz sampling rate, 16-bit PCM, and mono-channel.

CCC-VPR2C2005-1000: This corpus contains speech from 1,000 male speakers aged 18-23, each of whom was required to utter 40 Chinese sentences 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 recorded through a GSM mobile phone and the second time the utterance was recorded through a landline telephone.

For more details on these two data sets, please visit the homepage of the CCC and check their corresponding links on the "*Corpora*" page.

Although the participants were allowed to use the data set(s) they already had to develop their system(s), the CCC also provided them with development data, and all tests were performed on the evaluation data later provided by the CCC. All the wave files in the selected data sets are of 8 kHz sample rate, 16-bit precision, mono, linear PCM format (some of them were converted from different sample rates).

3.4 Development Data

Two development data sets were provided, one for text-independent tasks and one for text-dependent tasks.

(1) Development Data for Text-Independent Tasks

This data set is taken from CCC-VPR2C2005-1000. It contains data from 300 speakers randomly selected from the original 1,000 speakers. Data for each speaker includes 2 utterances, corresponding to one land-line (PSTN) channel utterance and one cellular-phone (GSM only) channel utterance. So the development data includes a total of 600 (=300×2) utterances.

Each utterance is divided into several segments, where there is at least 1 segment longer than 30 seconds, which can be used to train the speaker model. The other part is divided into several shorter segments, which can be used for testing. The order of the segments of different lengths in an utterance is determined randomly.

The relationships between the segment files and their speaker identities are defined in a key file shipped with the data set. This file also includes other necessary information, such as channel type and gender.

(2) Development Data for Text-Dependent Tasks

This data set is taken from CCC-VPR3C2005. It contains utterances partly selected from 5 male speakers' data and 5 female speakers' data. The data can be used as samples to listen or to perform some simple tests, but it is not sufficient to be used for clustering, for example, training channel-specific UBMs as the other data set can. In this data set, each speaker's data comes from three microphones, marked as micl, micr, and micu, respectively. For each channel, the data for each speaker includes 5 utterances repeated 4 times, as well as 21 other unrepeated utterances. The relationship between the segment files and their speaker identity are defined in a key file shipped with the data set. This file also includes other necessary information, including channel type and gender.

This data set also provides transcriptions for the training utterances, which can be accessed via the indexes listed in the key file.

3.5 Evaluation Data

The general features of the evaluation data, such as involved channel types and speaking styles, are the same as those of the development data. However, the speakers in the two stages' data sets do not overlap.

The training and trial lists were shipped with the evaluation data set, which covers the predefined evaluation tasks, i.e., combinations of text-independent or text-dependent, identification or verification, single-channel or cross-channel. For verification tasks, the ratio of testing samples for true-speakers and imposters is about 1:20; while for identification tasks, the ratio of testing samples for in-set (matched) and out-of-set (non-matched) cases is about 1:1.

The key files mapping test samples with their speaker identities were sent to the participants, along with the performance rankings and evaluation scripts, after all results were received and verified.

3.6 Participants

Eight research sites participated in the CCC 2006 SRE. The sites and their affiliations are:

- **NTUT-EE:** Speech Lab, Department of Electronic Engineering, National Taipei University of Technology, Taipei.
- **UVA-CS:** Computer Science Department, Universidad de Valladolid, Valladolid
- **CUHK-EE:** Department of Electronic Engineering, The Chinese University of Hong Kong, HKSAR.
- **THU-EE:** Department of Electronic Engineering, Tsinghua University, Beijing.

- **I2R-SDPG:** Speech and Dialogue Processing Group, Institute for Infocomm Research, Singapore.
- **EPITA:** BiOSECURE-EPiTA-FRiBOURG-GET, Le KREMLiN-BiCETRE
- **UT-ITS:** Institute of Technology and Science, The University of Tokushima, Tokushima
- **SINICA-IIS:** Institute of Information Science, Academia Sinica, Taipei.

3.7 Results

Although in total there were 6 tasks, no results for text-dependent single-channel speaker verification were submitted. A total of 17 systems from the eight participants were submitted for the remaining 5 tasks.

(1) Identification tasks

Only one test result for the text-independent cross-channel speaker identification task (abbreviated as *i-ti-c*) and two test results for the text-independent single-channel speaker identification task (abbreviated as *i-ti-s*) were submitted. The P_{IC}'s of these systems are shown in Table 2.

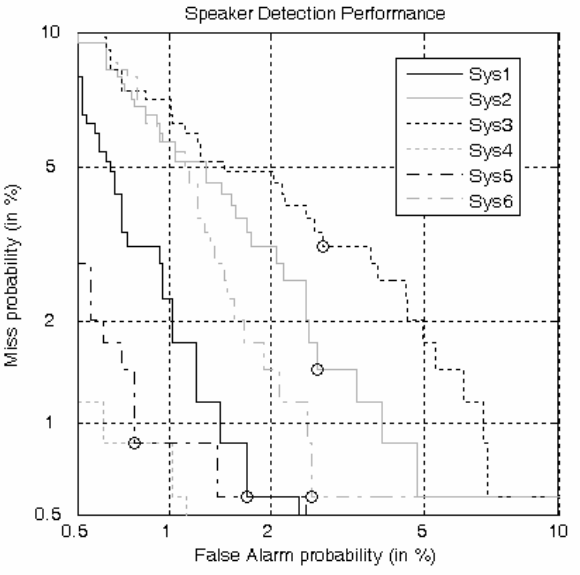

Fig. 1. DET curves for the *v-ti-s* task

(2) Verification tasks

The remaining 14 systems were for the verification tasks, particularly, 6 for the text-independent single-channel speaker verification task (abbreviated as *v-ti-s*), 7 for the

text-independent cross-channel speaker verification task (abbreviated as *v-ti-c*) and 1 for the text-dependent cross-channel speaker verification task (abbreviated as *v-td-c*).

Table 2. Identification test results

	i-ti-c	*i-ti-s*
Sys1	86.45%	
Sys2		99.33%
Sys3		97.16%

The DET curves and corresponding minimum C_{Det}s for the above tasks are given in Fig. 1 and Table 3, Fig. 2 and Table 4, Fig. 3 and Table 5, respectively. Note that the system IDs for each task are assigned independently.

Table 3. The minimum C_{Det}s for the systems in the *v-ti-s* task

System	Sys1	Sys2	Sys3	Sys4	Sys5	Sys6
C_{Det} (×100)	1.1	2.1	3.0	0.6	0.8	1.6

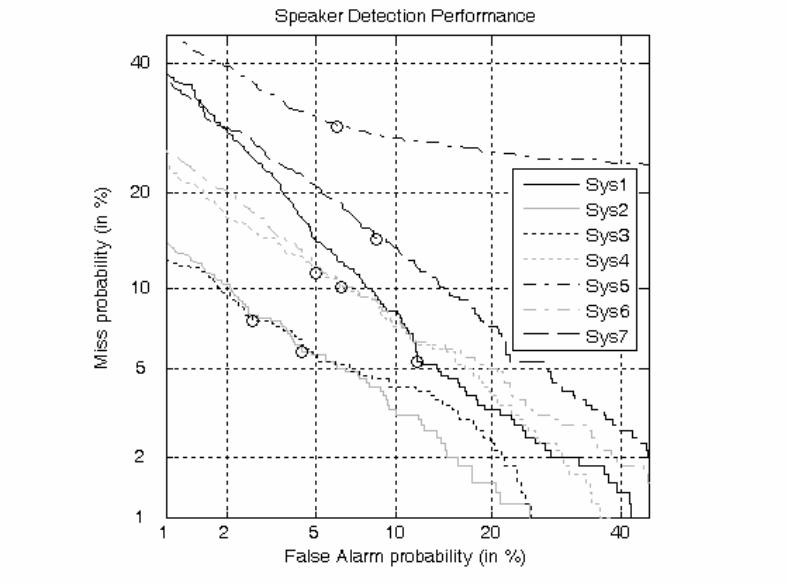

Fig. 2. DET curves for the *v-ti-c* task

Table 4. The minimum C_{Det}s for the systems in the *v-ti-c* task

System	Sys1	Sys2	Sys3	Sys4	Sys5	Sys6	Sys7
C_{Det} (×100)	8.6	5.1	5.2	8.2	17.8	8.2	11.5

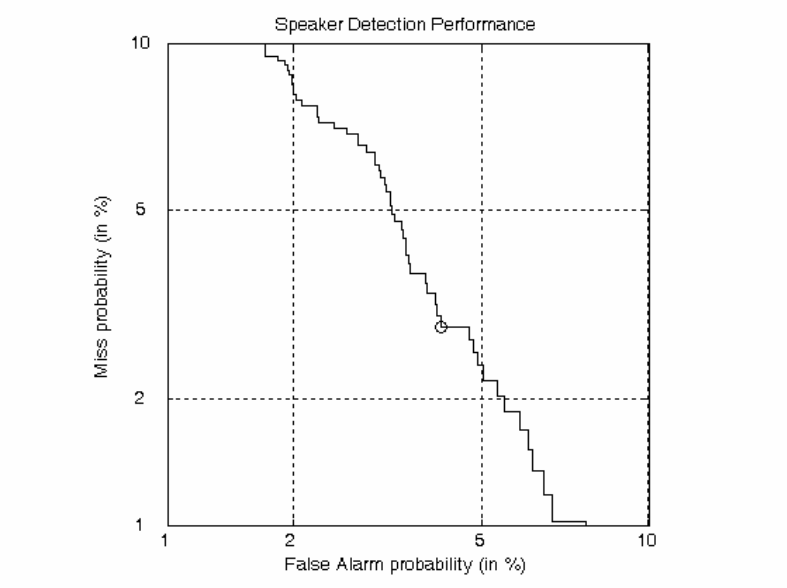

Fig. 3. DET curves for the *v-td-c* task

Table 5. The minimum C_{Det}s for the systems in the *v-td-c* task

System	Sys1
C_{Det} (×100)	3.53

As shown in the above results for the text-independent identification and verification tasks, the overall system performance in a cross-channel environment is worse than that in a single-channel environment, even though the cross-channel environment involves only two channel types, GSM and land-line. This phenomenon reveals that the channel effect is still a great impediment for speaker recognition. In light of this, the CCC is planning to collect corpora covering more complicated cross-channel environments, including various transmission channels and handsets.

4 Summary and Perspective

The CCC 2006 SRE began on Feb. 01, 2006 [8], and the conference presentation will be held on Dec. 16, 2006. Although this is the first time for this evaluation event to be carried out, the CCC would like to continuously support, improve and develop it into a series of events in the near future. This SRE was designed to be open to all, with announced schedules, written evaluation plans and follow-up workshops. The purpose of the evaluation is to provide additional chances for researchers and developers in this field to exchange their ideas and to help push forward, especially, corresponding work on Chinese language data. The CCC intends to use the experience gained this year in designing future evaluations. Any site or research group desiring to participate

in future evaluations is welcome, and should contact Dr. Thomas Fang Zheng (fzheng@tsinghua.edu.cn).

References

1. Campbell, J. P.: Speaker recognition: A tutorial. Proceedings of the IEEE. 85(9):1437-1462 (1997)
2. Reynolds, D. A.: An overview of automatic speaker recognition technology. In Proc. of ICASSP, 5: 4072-4075, Orlando, Florida, (2002)
3. Bimbot F., Bonastre J.-F., Fredouille C., Gravier G., Magrin-Chagnolleau I., Meignier S., Merlin T., Ortega-Garcia J., Petrovska-Delacretaz D., and Reynolds D.-A.: A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing, 4:430–451 (2004)
4. Zheng, T. F.: The Voiceprint Recognition Activities over China-Standardization and Resources. Oriental COCOSDA 2005, pp.54-58, December 6-8, Jakarta, Indonesia (2005)
5. Przybocki, M. A. and Martin, A. F.: NIST speaker recognition evaluations. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp. 331-335, Grenada, Spain, (1998).
6. Doddington, G.R., Przybycki, M.A., Martin, A.F., Reynolds, D.A.: The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective. Speech Commun. 31(2-3) (2000) pp.225-254.
7. Martin, A. and Przybocki, M.: The NIST Speaker Recognition Evaluations: 1996-2001. In Proc. 2001: A. Speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece, June 18-22, 2001, pp. 39-43.
8. http://www.iscslp2006.org/specialsessions.htm