Oriental COCOSDA: Past, Present and Future

Shuichi ITAHASHI*+, Chiu-yu TSENG, Satoshi NAKAMURA*****

* National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan,

+ National Institute of Informatics (NII), Tokyo, Japan

** Institute of Linguistics, Academia Sinica, Taipei, Taiwan

*** ATR Spoken Language Communication Research Laboratories (ATR-SLC), Kyoto, Japan

{s.itabashi@aist.go.jp, itabashi@nii.ac.jp}, cytling@sinica.edu.tw, satoshi.nakamura@atr.jp

Abstract

The purpose of Oriental COCOSDA is to exchange ideas, to share information and to discuss regional matters on creation, utilization, dissemination of spoken language corpora of oriental languages and also on the assessment methods of speech recognition/synthesis systems as well as to promote speech research on oriental languages. A series of International Workshop on East Asian Language Resources and Evaluation (EALREW) or Oriental COCOSDA Workshop has been held annually since the preparatory meeting held in 1997. After that, we have had a series of workshops every year in Japan, Taiwan, China, Korea, Thailand, Singapore, India and Indonesia. The Oriental COCOSDA is managed by a convener, three advisory members, and 21 representatives from ten regions in Oriental countries. We need much more Pan-Asia collaboration with research organizations and consortia, though there are some domestic activities in Oriental countries. We note that speech research has become popular gradually in Oriental countries including Malaysia, Vietnam, Xinjang Uygur Autonomous Region of China, etc. We plan to hold future Oriental COCOSDA meetings in these places in order to promote speech research there.

1. Introduction

It has been well understood that it is necessary to collect and maintain large amounts of speech data of various kinds, allowing unrestricted access so that they can be utilized for research and development as well as for recognizer performance assessment. Utilization of common speech corpora will increase repeatability and objectivity of speech research. From the linguistic or cultural viewpoint, it is necessary and important to preserve speech data of various languages, especially those that are becoming extinct. It is said that many local languages or dialects are disappearing by the day. Hence there is a pressing need to preserve natural record of such languages. This is another important purpose of speech databases. A collection of data to be used for this purpose is called a speech database or a speech corpus as shown in Fig. 1.

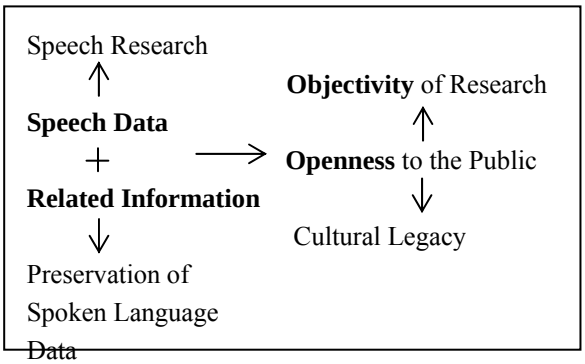

Fig. 1 Necessity of Speech Corpora

COCOSDA was established in 1991 to promote international cooperation in developing speech corpora and also in coordinating assessment methods of speech input/output systems (Campbell, 2000). It is an acronym of the International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment. It holds meetings every year as a satellite workshop of ICSLP and Eurospeech (later INTERSPEECH) conferences. EuroCOCOSDA was established as a suborganization in 1993 and later we started the Oriental COCOSDA as described below.

In the following, section 2 describes the history of Oriental COCOSDA briefly. Section 3 introduces organization of Oriental COCOSDA. Section 4 outlines the past annual meetings of Oriental COCOSDA, section 5 mentions the future plan, and section 6 concludes the paper.

2. Brief History

At the COCOSDA Workshop in Yokohama, Japan, in 1994, it was proposed by S. Itahashi (first author of the present paper) that East-Asian countries set up an organization to exchange ideas, to share information, and to discuss regional issues on spoken language processing. East Asian languages exhibit a wide range of characteristics which result in very different problems from European languages

(1) They embody considerable varieties arising from different language families;

(2) They use different orthographic systems, such as Chinese ideographic characters, Korean syllabic alphabet,

and Japanese alphabet;

(3) They use various systems of Romanization.

It is quite natural to suppose that there would be ways of processing these languages which are different from and more suitable than those adapted to European ones.

It had been recognized that it was necessary to create various kinds of speech and language corpora available for common use and to coordinate the system for utilization both in the process of research and development and in the performance evaluation of various speech systems. However, efforts to materialize this "recognition" were on a small scale and dispersive in oriental countries. We thought that it was a pressing need to coordinate these efforts from the viewpoints of not only the academic significance but also international cooperation of industry.

There were already several organizations in each oriental country but unfortunately with little or no mutual communication. Considering that, it was necessary at first, to prepare a common framework step by step to collect, create, store, distribute and share the speech and language data for the progress in future research on speech and language and on related fields of research. Researchers representing China, Korea, and Japan agreed to set up such an organization that coordinates problems related to speech and text corpora, speech recognition and synthesis, and speech input/output systems assessment methods; we had come to establish Oriental COCOSDA.

The purpose of Oriental COCOSDA is to exchange ideas, to share information and to discuss regional matters on creation, utilization, dissemination of spoken language corpora of oriental languages and also on the assessment methods of speech recognition/synthesis systems as well as to promote speech research on oriental languages. The Oriental COCOSDA Preparatory Meeting was held at the University of Hong Kong in 1997. After the preparatory meeting, we have had a series of workshops every year in Japan, Taiwan, China, Korea, Thailand, Singapore, India, and Indonesia (Itahashi, 2004).

3. Organization

The Oriental COCOSDA is managed by the convener, three advisory members from China, Japan and Korea, and 21 representatives from ten regions in Oriental countries including China, Hong Kong, India, Indonesia, Japan, Korea, Mongolia, Singapore, Taiwan, and Thailand. There are some domestic activities in Oriental countries. GSK (Linguistic Resources Association) was launched in Japan in 1999, SITEC (Speech Information Technology Industry Promotion Center) in Korea in 2001, and Chinese LDC in 2002. There is also CCC (Chinese Corpus Consortium) in China. We need much more collaboration with these organizations.

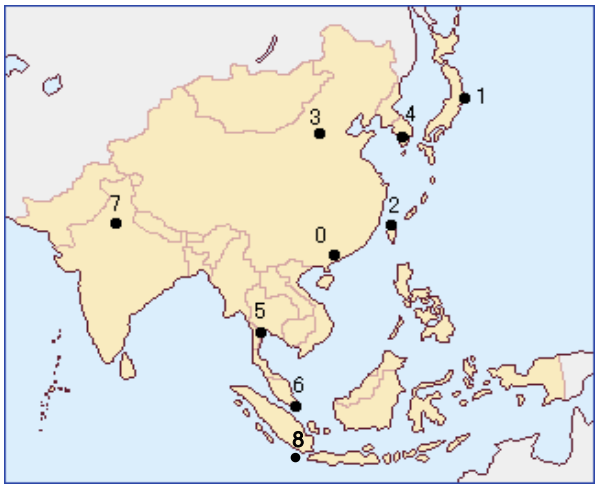

Fig. 2 Oriental COCOSDA Workshop sites

4. Outline of Annual Meetings

The Oriental COCOSDA Preparatory Meeting was held at the University of Hong Kong in March, 1997. At the Hong Kong meeting, Prof. H. Fujisaki, Professor Emeritus of University of Tokyo, delivered an overview of COCOSDA and pointed out general and regional problems on corpus studies. Prof. S. Itahashi proposed to hold the first workshop of Oriental COCOSDA in Tsukuba, Japan in May, 1998. Prof. Fujisaki clarified the definition of “Oriental” could be twofold, regional and linguistic (non-European). It was discussed that members of Oriental COCOSDA should be either those who live and work in oriental districts, speak and study oriental languages or those who are interested in oriental language corpora and speech input/output systems standardization. It is understood that Oriental COCOSDA is a sub-organization of COCOSDA in the sense that the members of the former attend the meeting of the latter to report and discuss their activities.

4.1. Tsukuba Meeting

The first meeting was held in Tsukuba, Japan in May, 1998 where 30 papers were presented. There were sessions on speech corpora, assessment, orthography and Romanization, and prosodic notation. The speech corpora session included corpora for synthesis, recognition,

dialogue and text. The assessment session included that of synthesis and recognition. There were 54 participants coming mostly from Japan and some from China, Korea, Taiwan and Thailand. Prof. H. Fujisaki presented an overview on "International Efforts towards Coordination and Standardization of Speech Databases and Speech Input/Output Assessment Methods" referring to the activities of COCOSDA, ELRA, and LDC. An invited lecture was given by Prof. H. Suzuki of University of Tsukuba "On the use of linguistic database for linguistic research." in which he mentioned the importance of database use in linguistics based on his experience in the early stage of the linguistic database development. The scale of the meeting was not so big, but it offered an opportunity for speech researchers of oriental languages the first significant step toward international sharing of speech corpora.

4.2. Taipei Meeting

The second meeting was held in Taipei in May, 1999. We had about 110 participants including 70 from Taiwan and 40 from overseas, and about 10 more of on-site registration. The participants were mostly from Taiwan and international attendants came from Japan, China, Korea, Thailand, U.S.A. and France. We had 36 presentations, 4 invited talks and a panel discussion. The invited talks were given by Prof. H. Fujisaki on Information retrieval based on human-machine dialogue, Dr. B-H Juang of Bell Labs. Lucent Technologies on Experiment Design for Speech Recognition and Understanding, Dr. Khalid Choukri of ELRA/ELDA on European Language Resources Association, and Prof. K. Shikano of Nara Institute of Science and Technology on Volunteer-Based IPA Japanese Dictation Free Software Project. The main discussion in the panel was devoted to coordination issues such as national, regional, and international initiatives and programs, and to defining the new cooperative trends within language resources and evaluation seeking the right model for East-Asia considering experiences in North America and Europe.

4.3. Beijing Meeting

The third meeting was held at Beijing International Convention Center in China on Mon. 16 Oct. just before ICSLP 2000. This time neither prior registration nor registration fee were necessary, which was different from the former two meetings. It was on a rather small scale, as the meeting was held adjacent to the parent COCOSDA meeting, but we had quite lively presentations and discussions with eight reports presented. Prof. S. Itahashi presented a brief overview of Oriental COCOSDA activities in the opening remarks. Speakers from China, Korea, Taiwan and Thailand reported the speech-related projects and the present status of spoken language corpora creation. Dr. K. Tanaka from Japan introduced JEIDA standard of symbols for TTS synthesizers. Mr. I. Dawa focused on the Mongolian language. Dr. C.-Y. Tseng from Taiwan made a report on labeling emphasis in Chinese speech. Prof. L.-S. Lee from Taiwan, then convener of COCOSDA, stated the new organization and future activity plan of COCOSDA. Prof. L. Du from China concluded the workshop.

4.4. Taejong Meeting

The fourth meeting was held in Taejon, Korea in August, 2001. It was a satellite event of ICSP (International Conference on Speech Processing) held also in Taejon. It was a one-day meeting and we had only 11 presentations of the reports, but we had participants not only from Korea but also from China, Japan, Thailand, Taiwan and Australia. It was hosted by the Speech Information Technology & Industry Promotion Center (SITEC) which was launched in May, 2001. The Center was supposed to be the Korean counterpart of LDC. It obtained similar amount of budget as LDC for its initial setup during the first five years; after that they are supposed to stand on their own feet and be self-supportive.

4.5. Hua Hin Meeting

The 5th meeting was held jointly with SNLP (Symposium on Natural Language Processing) in 2002 during May 9-11 in Hua Hin, Thailand. SNLP was initiated by NLP researchers in Thailand in 1993 and has been held biannually. There were about 100 participants mostly from Asia. We had five invited talks by four speakers: Prof. Emeritus H. Fujisaki of University of Tokyo on Information Retrieval and Modeling of Tonal Features of Speech. Prof. Fangxin Chen of IBM China Research Laboratory on Speech Synthesis for Tonal Languages, Prof. T. Tokunaga of Tokyo Institute of Technology on Natural Language Understanding and Action Control, and Prof. D. Yarowsky of Johns Hopkins University on Cross-Language Projection of Linguistic Knowledge. There were about 57 oral presentations including 28 regular papers, 17 short papers, 8 COCOSDA papers and 4 student papers. Among them, 23 were from Thailand, 14 from Japan, 5 from China, 3 from Korea and India, 2 from Taiwan and one from Malaysia, Indonesia and Guam except the student papers which were all from Thailand. General presentations were made in parallel for

NLP and speech processing using two rooms.

4.6. Singapore Meeting

The 6th meeting was also held jointly with PACLIC (Pacific Asia Conference on Language, Information and Computation) in Singapore in Oct. 2003 and 28 papers were presented. The participants were from Singapore, China, Taiwan, India, Indonesia, Japan, and Korea. There were two invited presentations: Prof. S. Itahashi of University of Tsukuba, Japan presented an "Overview of the East-Asian Activities on Speech Corpora and Assessment," and Prof. Fang Zheng of Tsinghua University, China talked on "Making Full Use of Chinese Speech Corpora." There were three sessions on Speech Input and Output, two sessions on Speech Corpora and the sessions of Assessment and Phonetic Systems of Oriental Languages of one session each.

4.7. Delhi Meeting

The 7th meeting was held in Delhi, India in Nov. 2004 together with iSTEPS (International Symposium on Speech Technology and Processing Systems) and also iSTRANS (International Symposium on Machine Translation, NLP and TSS). There were 13 invited talks and 53 presentations of speech-related papers and attended by over 150 participants coming mostly from all over India. International participants also came from Australia, France, China, Indonesia, Japan, Korea, Singapore, Taiwan, and U.S.A. In addition to oral presentations by authors, there were lively discussions throughout the meeting. The event also drew considerable local media coverage.

4.8. Jakarta Meeting

The 8th meeting was held in Jakarta in Dec. 2005. Initially, the workshop was planned to be held in Bali, but it was moved to Jakarta because of the bomb affair in Bali in October. There were 65 participants, mostly from Asia, and 22 presentations plus two invited talks. Among them, 9 were from Japan, 3 from China and Malaysia, 2 from Taiwan and Indonesia, and one from Korea, Thailand, Hong Kong, Singapore, and Mongolia. There was one session in the morning, and two sessions in the afternoon. Among them, two sessions were for speech corpora, and one session for speech recognition, speech synthesis, speaker identification and spoken dialogue. On the third day, there was a city sightseeing tour. This time we had participation from Malaysia for the first time. The invited talks are as follows: Corpus and technologies for ATR speech-to-speech translation by Dr. Satoshi Nakamura of ATR, Japan and Characteristics of Indonesian Language from Perspective of Language Technologies by Dr. A. A. Arman of ITB, Indonesia.

5. Future Plans

The 9th meeting is planned to be held in Malaysia in December 2006. Speech research has become popular gradually in Oriental countries including Vietnam, Xinjang Uygur Autonomous Region of China, etc. We are going to hold the Oriental COCOSDA meeting in these places in order to promote speech research there. We are also looking for other related conferences for possible joint sessions and/or special sessions to promote Oriental COCOSDA and its missions to research communities on speech and spoken language processing.

6. Conclusion

This paper has introduced the development of Oriental COCOSDA. LDC in U.S.A. and ELRA in Europe have contributed to the creation, collection, and distribution of speech corpora. GSK, SITEC and the Chinese LDC are expected to play the role of LDC or ELRA in Asia. Continuous speech corpora, especially those containing spontaneous speech of Asian languages, are still in preparation. In the future, it will be increasingly necessary to enrich speech and language corpora, especially to promote the collection and utilization of Asian language corpora, of which either tones or pitch accents are distinct phonetic features for some and a large number of them use non-alphabetic writing systems.

A recent milestone is the fact that the National Institute of Informatics in Japan has decided to start its activities on speech corpora together with GSK. This act will strengthen corpora resource sharing and standardization activities not only in Japan but also in East Asia. We therefore believe Oriental COCOSDA will continue to contribute to its cause and missions and play an ever more important role in the region.

For more information please refer to the following URLs.
http://www.slc.atr.jp/o-cocosda/
http://www.cocosda.org/

7. References

Campbell, N. (2000). COCOSDA – a Progress Report. *Proc. LREC 2000*, Athens, Greece, pp. 73-76.

Itahashi, S. (2004). Overview of the Asian Activities on Speech Corpora and Standardization. Invited paper, *Proc. iSTRANS-2004 and Oriental COCOSDA 2004*, Delhi, India, pp. 3-11.