The 2008 Oriental COCOSDA Book Project –
In Commemoration of the First Decade of Sustained Activities in Asia

Shuichi ITAHASHI*!, Chiu-yu Tseng+
* National Institute of Informatics (NII), Tokyo, Japan
! National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan
+ Academia Sinica, Taipei, Taiwan
E-mail: itabashi@nii.ac.jp, cytling@sinica.edu.tw

Abstract

The purpose of Oriental COCOSDA is to provide the Asian community a platform 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. Since its preparatory meeting in Hong Kong in 1997, annual workshops have been organized and held in Japan, Taiwan, China, Korea, Thailand, Singapore, India, Indonesia, Malaysia, and Vietnam from 1998. The organization is managed by a convener, three advisory members, and 26 committee members from 13 regions in Oriental area. In order to commemorate 10 years of continued activities, the members have decided to publish a book which covers a wide range of speech research. Special focus will be on speech resources or speech corpora in Oriental countries and standardization of speech input/output systems performance evaluation methods on which key technologies for speech systems development are based. The book will also include linguistic outlines of oriental languages, annotation, labeling, and software tools for speech processing.

1. Introduction

COCOSDA, an acronym of the International Committee for Coordination and Standardization of Speech Databases, was established in 1991 to promote international cooperation in developing speech corpora and coordinating assessment methods of speech input/output systems (Campbell, 2000). In 1994 it was proposed that a sub-organization for the Oriental community should be established to share linguistic features unique to the region. After a preparatory meeting held by interested members in Hong Kong in 1997, annual meetings have been held since 1998. The community has enjoyed increasing participation from the community and enthusiastic interests to organize future meetings, thus ensuring promising prospect of sustained activities in the future.

The Oriental COCOSDA Preparatory Meeting was held at the University of Hong Kong in March, 1997, where Prof. H. Fujisaki pointed out that 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 assessment. 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. So far, the Oriental COCOSDA workshops have been held at various places in Oriental countries since the first meeting held in 1998 in Tsukuba, Japan.

In order to commemorate 10 years of continued activities, the members have decided to publish a book which covers a wide range of speech research. Special focus will be on speech resources or speech corpora in Oriental countries and standardization of speech input/output systems performance evaluation methods on which key technologies for speech systems development are based. The book will also include linguistic outlines of oriental languages, annotation, labeling, and software tools for speech processing.

In the following, section 2 introduces a brief history of Oriental COCOSDA, section 3 discusses the organization of Oriental COCOSDA, section 4 outlines the features of Asian languages, section 5 shows the contents of the book, and section 6 is the conclusion.

2. Brief History of Oriental COCOSDA and Future Events

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. Typologically, East Asian languages exhibit a wide range of characteristics which result in very different problems from European languages. It is quite natural to assume that there would be different and more suitable ways of processing these languages other than those adapted from 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. There were already several organizations related to speech corpora in each oriental country but unfortunately with little or no mutual communication. 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 (Itahashi, 2004). Thus it was decided that 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. After the preparatory meeting held in Hong Kong in 1997, a series of annual workshops has been held in Japan, Taiwan, China, Korea, Thailand, Singapore, India, Indonesia, and Malaysia (Itahashi, Tseng & Nakamura, 2006; Tseng & Itahashi, 2007). Figure 1 shows the number of papers presented and the number of participants of the past 10 meetings.

0. 1997 Hong Kong (11 participants)
1. 1998 Tsukuba, Japan (30 papers, 54 participants)
2. 1999 Taipei, Taiwan (44, 120)
3. 2000 Beijing, China (8, 20)
4. 2001 Taejon, Korea (11, 25)
5. 2002 Hua Hin, Thailand (24, 96) w/SNLP
6. 2003 Sentosa, Singapore (28, 60) w/ PACLIC
7. 2004 Delhi, India (55, 150) w/SPLASH
8. 2005 Jakarta, Indonesia (24, 65)
9. 2006 Penang, Malaysia (31, 60)
10. 2007 Hanoi, Vietnam (34, 75)

Fig. 1 Number of papers and participants in the meetings

The 5th, 6th and 7th meetings were joint events with other related conferences. The size of these meetings have reached at over 30 papers and 60 plus participants since 2005, not very large by international standard, but they did provide the only occasion where the members met face-to-face annually to discuss Oriental COCOSDA business. In addition to paper presentation, the successive conveners have also set up a tradition to hold a business meeting that consists of Oriental COCOSDA activity reports, country reports and announcement of future conferences, thus bringing more countries and regions to join the meetings over time. Figure 2 shows the participating regions in each year.

0 China, Japan, Korea, Taiwan (CJKTw), Hong Kong (HK)
1. CJKTw
2. CJKTw, Thailand (Th), France (F), U.S.A.
3. CJKTw, Th, Mongolia (Mg)
4. CJKTw, Th, Australia (Au)
5. CJKTw, Th, India (Id), Indonesia (Is), Guam
6. CJKTw, Th, Id, Is, Singapore (S)
7. CJKTw, Id, Is, S, Au, F, U.S.A.
8. CJKTw, Th, Is, Malaysia, Mg, HK
9. CJKTw, HK, Id, Is, My, Nepal, S, Vietnam (V), Yemen
10. CJKTw, HK, Id, Is, Pakistan, S, Th, U.S.A., V

Fig. 2 Region breakdown of the meeting participants

The outline of the past meetings including the preparatory meeting held in 1997 in Hong Kong was mentioned in (Tseng & Itahashi, 2007) up to the 8th meeting (2005). Sections 2.1 and 2.2 below are brief update of the latest two meetings.

2.1 Oriental COCOSDA 2006 Workshop, Penang, Malaysia

The 9th meeting was held in December 9-11, 2006 in Penang, Malaysia and drew over 60 participants from Malaysia, China, India, Indonesia, Japan, Nepal, Taiwan, Thailand, Vietnam, Yemen and Jordan. It was during this meeting that participation from Nepal and Vietnam came for the first time. There were three invited keynote speeches and 31 oral presentations. The invited speeches included 1) SITEC creation and distribution of language resources in Korea, 2) Corpus-based synthesis of fundamental frequency contours using generation process model and automatic preparation of training corpora (Japan), and 3) standardization of speech corpora for Indian languages (India). Oral sessions cover topics on corpus and technologies, emotion and speech recognition, speech synthesis, phonetics and language teaching. A special session was created to include an update on a new cross-country APEC project A-Star of speech to speech translation. There were 7 papers from Malaysia, 4 papers from China, Japan, Taiwan, and Vietnam, 3 papers from India, one paper from Indonesia, Nepal, Thailand and Yemen.

One feature of this meeting was how Oriental COCOSDA established closer relationship with the community of SIG-CSLP (Chinese Spoken Language Processing), a special interest group under ISCA (International Speech Communication Association). ISCSLP 2007, an International Symposium on Chinese Spoken Language Processing, a biannual event since 1996, was held from December 13th the same year at Singapore. Participants of Oriental COCOSDA were invited to present related works to a bigger and largely Chinese community at reduced fees. Two special sessions were organized to accommodate 12 papers including research on Japanese, Indian and Vietnamese other than Chinese. The interaction proved to be very positive and mutually beneficial to both communities; the CSLP community showed enthusiastic interests in neighboring languages and speech related research; both communities looked forward to more future interaction.

2.2 Oriental COCOSDA 2007 Workshop, Hanoi, Vietnam

The 10th meeting was held in December 4-6 in Hanoi, Vietnam and drew over 70 participants from 10 countries of Asia and U.S.A. The three invited talks were on 1) Computer supported human-human multilingual communication (USA and Germany), 2) Network-based Asian-Pacific speech-to-speech translation (Japan), and 3) an introduction to spoken language recognition (Singapore). There were 10 papers from Malaysia, 5 papers

from India, Japan, Taiwan, and Vietnam, one paper from Hong Kong, Indonesia, Korea, and Nepal. We had the first participation from Pakistan this year. A total of 31 oral papers covering areas of speech segmentation and annotation tool, speech prosody and labeling, speech recognition and synthesis, corpus technology and speech processing, models and systems were presented.

In summary, a decade since its first meeting in 1998, 10 annual Oriental COCOSDA workshops have been held. Participation of countries (by country report) increased from the initial 5 to the current 13. The meeting has also reached the state of attracting over 60 participants when held independently from other events.

2.3 Future Oriental COCOSDA Meetings

A general meeting was held on December 5th 2007 during the 10th n Oriental COCOSDA meeting at Hanoi, Vietnam. Among the resolutions are two future Oriental COCOSDA meetings. The 11th meeting will be organized by ATR, Japan and held in November 25-27, 2008 in the Kyoto vicinity; the 12th meeting will be organized by Xinjiang University, China and held in August 2009 at Urumuqi, Xinjiang. One feature of the 2009 meeting was to follow the 2006 meeting and work neck to neck with ISCSLP 2008 at the same location. We believe more regular interaction with the CSLP community can be expected in the future.

3. Organization

The Oriental COCOSDA is managed by the convener, three advisory members from China, Japan and Korea, and 26 representatives from 13 regions in Oriental countries including China, Hong Kong, India, Indonesia, Japan, Korea, Malaysia, Mongolia, Nepal, Singapore, Taiwan, Thailand, and Vietnam.

Related domestic activities in some of the member countries also took place in the past decade. Linguistic Resource Association (GSK) was launched in Japan in 1999, Speech Information Technology Industry Promotion Center (SITEC) in Korea in 2001, and Chinese LDC in 2002. There is also Chinese Corpus Consortium (CCC) in China. The Speech Resources Consortium (SRC) of National Institute of Informatics started its activities in 2006 in Japan. We believe much more collaboration with and among these organizations are needed.

4. Features of Oriental Languages

Asia is a multilingual region and we notice that the diversity of languages is larger than Europe, and yet typologically Asian languages form a class of its own. Speech researches were emerging; speech corpora were required; cooperation among countries was necessary; organizations for speech corpora were needed. It became evident that by now most of these countries have all launched large scale projects on speech corpora collection with different focuses that reflect the linguistic diversities and properties of the region even further. For example, though both China and India are large countries, their tasks have been very different. China has one official language (Putonghua) and one writing system (the Chinese ideograph). India, on the other hand, has over a dozen of official languages and writing systems. Their main efforts have been standardizing multi-lingual speech corpora in India and speech-to-speech translation. Indonesia and Malaysia have similar multi-lingual multi-script problems as India, but Indonesia asks for more concrete actions towards platform standardization and resource exchange while Malaysia's efforts on speech corpora are still efforts in the hands of a very small group academic professions.
There are a variety of language families of Asian languages according to Ethonologue.com:

1. Austronesian (1268 languages): Malay, Indonesian, etc.
2. Sino-Tibetan (403): Chinese, Tibetan, Burmese, etc.
3. Austro-Asiatic (169): Khmer, Vietnamese, etc.
4. Tai-Kadai (76): Thai, Lao, etc.
5. Dravidian (73): Tamil, Telugu, etc.
6. Altaic (66): Mongolian, Turkic, Korean, etc.
7. Japanese (12): Japanese, Ryukyuan, etc.

cf. Indo-European (449)

We can also see some varieties in letters, tone and word order in Asian languages:

1. Proper letters: Burmese, Chinese, Japanese, Khmer, Korean, Thai, etc.
2. Latin letters: Indonesian, Malay, Vietnamese, etc.
3. Tonal languages: Burmese, Chinese, Lao, Thai, Vietnamese, etc.
4. Word order: SOV, SVO, VSO, VOS (S: subject; V: verb; O: object).

Regarding word boundary in text, there are two varieties:

1. No space between words: Burmese, Chinese, Japanese, Khmer, Lao, Thai, etc.
2. Space between words: Indonesian, Malay, Mongolian, Vietnamese, etc.

In addition, there are non-unique Romanization systems for languages using non-Latin letters.

According to the development of information and communication technology, we are at a stage where we are now able to cope with automatic processing of oriental languages which had not been possible before. So the time has come for the community to integrate a decade of activities and document collective efforts by publishing a book on speech corpora and speech processing of Oriental languages.

5. Oriental COCOSDA Book Project

The title of the book is "Resources and Standards of

Spoken Language Systems – Advances in Oriental Spoken Language Processing–" edited by Shuichi Itahashi and Chiu-yu Tseng, authors of the present paper. The book will consist of 9 chapters and an appendix as shown below.

1. Introduction
This chapter discusses the mission of Oriental COCOSDA, etc.
2. Outline of Oriental languages
This chapter introduces phonology, phonetics, prosody and orthography of Chinese, Indian languages, Indonesian, Japanese, Korean, Malay, Mongolian, Nepali and Vietnamese.
3. Data centers and corpora
This chapter outlines the organizations and activities of several organizations handling speech corpora as shown below.
3.1 Language Resource Association (GSK) in Japan.
3.2 Speech Information Technology & Industry Promotion Center (SITEC) in Korea.
3.3 Chinese Linguistic Data Consortium.
3.4 Chinese Corpus Consortium (CCC).
3.5 NII-Speech Resources Consortium (NII-SRC) in Japan.
3.6 ATR-NICT speech and text corpora in Japan.
3.7 Introduction to Korterm in Korea.
3.8 Association for Computational Linguistics and Chinese Language Processing (ACLCLP) in Taiwan.
4. Speech corpora of Oriental Languages
This chapter describes corpora by speech type such as monologue, dialogue or conversation, meeting speech, in-car speech, language learners' speech, emotional or expressive speech, prosodic corpora, multilingual corpora, simultaneous interpretation corpora, multimodal corpora, sound and noise database, and code-switching.
5. Performance evaluation of synthesizers and recognizers
This chapter recapitulates the standards of performance evaluation methods of speech synthesizers and recognizers, standard of symbols for speech synthesizers and recognizers, and the method of performance comparison.
6. Annotation and labeling
This chapter summarizes phoneme labeling, prosody labeling, emotion labeling, segmentation and alignment.
7. Software tools
This chapter describes various software tools for speech processing, speech recognition, speech synthesis, talking head, prosody model, and comparison of speech tools.
8. Orthographic transcription and Romanization system
This chapter treats orthographic system and Romanization of Japanese, Korean, and Thai. It also mentions grapheme to phoneme conversion of Hindi, standard Malay, and Mongolian.
9. Conclusion
This chapter summarizes the activities of Oriental COCOSDA and states the future plans.
Appendix: Historical description of Oriental COCOSDA including conveners, representatives, workshops, and photos, etc.

6. Conclusion

This paper summarizes a decade of sustained activities of Oriental COCOSDA including its history, organization, related initiatives in Oriental countries, as well as unique linguistic features. The regional language specific features and ever-growing necessity to develop speech corpora and speech related research appear to be the drive force of Oriental COCOSDA community to remain in close contact through contributing and organizing annual meetings. In addition to speech, non-alphabetic writing systems will also continue to be another research focus. A book project to commemorate 10 years of Oriental COCOSDA is another highlight. The themes and content of the book signifies ten years of collective efforts as well as a momentum of the research community that continues to grow. Though many books on speech research have been published, none in existence covers the processing of various Oriental languages as comprehensively as the one described in this paper. This book will serve as the first comprehensive introduction to speech processing, speech corpora and standardization activities of Oriental languages. We believe it will be useful not only to researchers in Oriental regions and languages, but also to speech research communities world-wide. Further information on COCOSDA and Oriental COCOSDA are available at the following URLs.
http://www.slt.atr.co.jp/o-cocosda/
http://www.cocosda.org/

7. Acknowledgements

We would like to express our sincere gratitude to all past organizers of annual meetings, all the committee members, and most of the steadily increasing Oriental COCOSDA research community.

8. References

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

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

Itahashi, S., Tseng, C-Y, Nakamura, S. (2006). Oriental COCOSDA: Past, Present and Future. In *Proceedings of LREC 2006*, Genoa, Italy, pp. 753--756.

Tseng, C-Y, Itahashi, S. (2007). Oriental COCOSDA – Language Resource Efforts of the Greater Asian Region. In *Proceedings of First International Symposium on Universal Communication (ISUC 2007)*, Kyoto, Japan, pp. 118--121.