Oriental-COCOSDA—

Language Resource Efforts of the Greater Asian Region

Chiu-yu TSENG[1] and Shuichi ITAHASHI[2]

[1]*Institute of Linguistics, Academia Sinica, Taipei, Taiwan*
[2]*National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan,*
[2]*National Institute of Informatics (NII), Tokyo, Japan*

Abstract This paper reports the O-COCOSDA community's past activities to promote speech research and spoken language corpora, and its future outlooks, focusing on language dependent features unique to the region.

1. Introduction

In 1991COCOSDA (the International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment) was established to promote international cooperation in developing speech corpora and also in coordinating assessment methods of speech input/output systems [1]. EuroCOCOSDA was established as a sub-organization in 1993 and the Oriental COCOSDA was established soon afterwards. Detailed reports of O-COCOSDA history and activities have been documented [2, 3]. This paper summarizes past reports and updates recent activities using materials from the last O-COCOSDA workshop held in Penang, Malaysia in December, 2006.

The need to collect and keep 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 has been recognized for decades. Utilization of common speech corpora will also increase repeatability and objectivity of speech research [3], something classic field phonetic research is unable to offer. The need to collect speech corpora is more pressing in Asia for the following reasons: (1.) Some of the oldest languages of the world remain but facing distinction; (2.) many languages writing system; and (3.) many countries are multi-lingual (India, Indonesia, Malaysia); (4.) the majority writing systems in terms of user population are either ideographic-syllabic (Chinese), alphabetic plus ideographic (Japanese), syllabic alphabetic (Korea) instead of alphabetic; and (5.) the various systems used for Romanization. In addition, the majority of East Asian languages are tone languages. In short, Asian languages and speeches possess unique characteristics both linguistically and culturally; put forth inimitable problems for technology development; and assuming the responsibility of cultural heritage documentation preservation as well. All of the above sets the region aside from Western and European languages.

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 updates activities from the O-COCOSDA 2006 country reports, section 6 on O-COCOSDA's activities outside Asia, Section 7 reports future plans, and section 8 concludes the paper.

2. Brief History [2, 3]

At the COCOSDA Workshop in Yokohama, Japan, in 1994, Professor Shuichi Itahahsi from Japan proposed that East-Asian countries set up an organization to exchange ideas, share information, and discuss regional issues on speech data collection and spoken language processing. As mentioned in Section 1, languages in East Asia exhibit a wide range of characteristics unlike European languages, it was quite natural to assume that instead of adopting European ways to process these languages, there should be more suitable ways developed for these languages only. 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.

At that time, most of the developed and developing East Asian countries have began such efforts through various organizations, but unfortunately with little or no mutual communication. The proposal was immediately well received by researchers representing China, Korea, and Japan to a common framework required 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. Therefore, an organization that coordinates problems related to speech and text corpora, speech recognition and synthesis, and speech input/output systems assessment methods was established, namely, the Oriental COCOSDA. Its purpose is to exchange ideas, 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, Indonesia, and Malaysia. The number of countries that joined the organization has also increased from the initial 5 to the current 13.

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 by the end of 2006, including China, Hong Kong, India, Indonesia, Japan, Korea, Malaysia, Mongolia, Nepal, Singapore, Taiwan, Thailand and Viet Nam. Professor Shuichi Itahashi of Japan was the first convener and served in that capacity until the end of 2005. Dr. Chiu-yu Tseng of Taiwan succeeded him from January, 2006.

Since the establishment of O-COCOSDA, there have been also been related domestic activities in Oriental countries, for example, GSK (Linguistic Resources Association) was launched in Japan in 1999, SITEC (Speech Information Technology Industry Promotion Center) in Korea in 2001, CCC (Chinese Corpus Consortium) in China and later the Chinese LDC in 2002. O-COCOSDA is on good terms with these organizations and plans much more collaboration with them in the future. [3]

So far all members are voluntary and there is membership fees involved. Annual meetings have been held since 1998 in various countries or region, with the understanding that the local organizers are solely responsible for funds to host these meetings.

4. Annual Meetings

The Oriental COCOSDA Preparatory Meeting was held at the University of Hong Kong in March, 1997. 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. “Oriental” was defined by Professor Fujisaki in two ways, 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. [2, 3]

4.1. O-COCOSDA 1998 Workshop, Tsukuba, Japan

The first meeting was held in Tsukuba, Japan in May, 1998, and drew 54 participants. Though the majority of participants were from Japan, there were also participations from China, Korea, Taiwan and Thailand. There were 2 invited papers and 30 oral papers. The invited papers addressed international precedence including COCOSDA, ELRA and LDC, and the use of databases for linguistic research. Papers covered topics on (1.) speech corpora for synthesis, recognition, dialogue and text; (2.) assessment; (3.) orthography and Romanization; and (4.) prosodic notation. Though it was not a large scale meeting, it drew undivided attention from the participants, laid a solid foundation for future workshops. It was agreed that the meeting should be held annually, and the international participants

agreeably set up the order to host future meetings.

4.2. O-COCOSDA 1999 Workshop, Taipei, Taiwan

The second meeting was held in Taipei in May, 1999 that drew over 120 participants, including 70 from Taiwan and 40 from overseas, and about 10 more of on-site registration. International attendants came from Japan, China, Korea, Thailand, U.S.A. and France. There were 36 presentations, 4 invited talks and a panel discussion. Invited speeches addressed issues on information retrieval based on human-machine dialogue, speech recognition and understanding, European language recourses and association, and volunteered-based IPA Japanese dictation free software. Technical sessions covered topics on (1.) corpora for spoken language (3 sessions), (2.) assessment of speech recognition and synthesis, (3.) phonetics, (4.) models and systems and (5.) topics on corpora. 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. This workshop is the largest meeting to date that devoted solely on O-COCOSDA.

4.3. O-COCOSDA 2000 Workshop, Beijing, China

The third meeting was held at Beijing International Convention Center in China on Mon. 16 Oct, 2000, as a satellite event of ICSLP 2000. It was held adjacent to the parent COCOSDA meeting and without registration requirements. The meeting was on a rather small scale of only eight presentations, with speakers from China, Japan, Korea, Taiwan and Thailand, but discussions were vigorous. Convener Professor S. Itahashi presented a brief overview of Oriental COCOSDA activities in the opening remarks, while presentations covered speech-related projects and the present status of spoken language corpora creation, JEIDA standard of symbols for TTS synthesizers, corpora on Mongolian language and prosodic labeling. Professor L.-S. Lee from Taiwan, then convener of COCOSDA, also attended the meeting and stated presented both new organization and future activity plans of COCOSDA.

4.4. O-COCOSDA 2001 Workshop, Taejong, Korea

The fourth meeting was held in Taejon, Korea in August, 2001 as a satellite event of ICSP (International Conference on Speech Processing), and was hosted by the Speech Information Technology & Industry Promotion Center (SITEC) which was newly launched in May, 2001. The meeting was a one-day meeting with only 11 presentations, but drew participants from China, Japan, Korea, Thailand, Taiwan and Australia. Topics included spoken language resources, speech corpora, speech technology development, speech tool comparison, tool evaluation and assessment tools. SITEC was modeled as the Korean counterpart of LDC by the Korean government, obtained similar amount of budget as LDC for initial setup during the first five years; and was expected to become self-supportive afterwards.

4.5. O-COCOSDA 2002 Workshop, Hua Hin, Thailand

The 5th meeting was held jointly with SNLP (Symposium on Natural Language Processing) in 2002 during May 9-11 in Hua Hin, Thailand and drew around 100 participants collectively. SNLP was initiated in 1993 by NLP researchers in Thailand and has been held biannually, most of them from Asia. There were five invited talks on information retrieval, modeling of tonal features of speech, synthesis for tonal languages, natural language understanding and action control, and 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 1 from Malaysia, Indonesia and Guam each. Student papers were all from Thailand. The meeting was held in parallel sessions with SNLP. It was the first time participants came from India, Indonesia and Malaysia.

4.6. O-COCOSDA 2003 Workshop, Singapore

The 6th meeting was also held jointly with PACLIC (Pacific Asia Conference on Language, Information and Computation) in Singapore in Oct. 2003. A total of 2 invited speeches and 28 papers were presented. The participants were from Singapore, China, Taiwan, India, Indonesia, Japan,

and Korea. There two invited speeches included an overview of the East-Asia activities on speech corpora and assessment by convener Professor S. Itahashi and Chinese speech corpora. Paper presentations covered topics on (1.) speech input and output (3 sessions), (2.) speech corpora (2 sessions), (3.) assessment (1 session) and (4.) phonetic systems of Oriental languages (1 session).

4.7. O-COCOSDA 2004 Workshop, Delhi, India

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. O-COCOSDA 2005 Workshop, Jakarta, Indonesia

The 8th meeting was held in Jakarta in Dec. 2005. Initially, the workshop was planned to be held in Bali, but due to terrorist bombing in October the organizers moved the meeting to Jakarta. In spite of the last minute change, the meeting still drew 65 participants, most of them from Asia. There were two invited talks and 22 presentations. 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 were two invited talks: Dr. Satoshi Nakamura of ATR, Japan on Corpus and Technologies for ATR Speech-to-Speech translation and Dr. A. A. Arman of ITB, Indonesia on Characteristics of Indonesian Language from Perspective of Language Technologies. During the 2-day 6-session meeting, 2 sessions were devoted to speech corpora and 1 session for speech recognition, speech synthesis, speaker identification and spoken dialogue each. It was during this meeting that participation from Malaysia came for the first time.

4.9. O-COCOSDA 2006 Workshop, Penang, Malaysia

The 9th meeting was held in December 9-11, 2007 in Penang, Malaysia and drew over 60 participants from Malysia, China, India, Indonesia, Japan, Nepal, Taiwan, Thailand, Vietnam, Yemen and Jordan. There were three invited keynote speeches and 31 oral presentations. The invited speeches included 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, and (3.) standardization of speech corpora for Indian languages. The oral sessions were on corpus and technologies, emotion and speech recognition, speech synthesis, phonetics and language teaching. A special session was set aside for a cross-country APEC project A-Star as well focusing on speech to speech translation. It was during this meeting that participation from Nepal and Viet Nam came for the first time. There were 4 papers from China and brought forth issues of code-mixing (with English) as well as semantic and phonetic topics; 3 papers from India on ELDA standard based Hindi speech corpora, Marathi speech database and grapheme to phone conversion for Hindi; 1 paper from Indonesian on speech-to-speech translation; 4 papers from Japan on corpus-based speech-to-speech translation systems, NII speech resources consortium and emotion labeling for automatic anger estimation; 7 papers from Malaysia on Malay speech modeling, automatic sub-word segmentation for Malay continuous speech, expressive text reader automation layer eXTRA, Malay speech synthesis, Malay phonetics, and Kelantanese developmental phonology; 1 paper from Nepal on writing based syllabification in Nepali; 4 papers from Taiwan on corpus phonetics, code switching, tone sandhi and speech technology; 1 paper from Thailand on NuThai Thai speech corpus; 4 papers from Vietnam on Vietnamese POS tagging, English-Vietnamese NP extraction and phrase-based English -Vietnamese machine translation, and key phrase extraction for information retrieval; and 1 paper from Yemen on Mehri Qishn root morphology. Nepal and Yeman participated the meeting for the first time.

In summary, there have been 9 annual O-COCOSDA workshops since its first meeting in 1998. Participation of countries (by country report)

increased from the initial 5 to the current 13, and has reached the state of attracting around 60 participants when held independent from other events.

5. National/Regional Activities in 2006

These meetings have been the only time where members met face to face to discuss O-COCOSDA business. Former convener Professor Shuichi Itahashi has set up a tradition to hold a business meeting that consists of O-COCOSDA activity reports, country reports and announcement of future conferences. At the 2006 business meeting, there was also a COCOSDA report in addition to O-COCOSDA country reports by COCOSDA Secretary Nick Campbell on (1.) resumed COCOSDA activities since 2005's meeting of the representatives at Interspeech, Lisbon; (2.) new COCOSDA convener Dafydd Gibbon; (3.) upcoming establishment of African-COCOSDA and (4.) a new COCOSDA Handbook proposal.

For the first time, the numbers of O-COCOSDA countries and regions have increased to 13 by 2006, of which 11 reports were presented at the business meeting in alphabetic order. They are: China, India, Indonesian, Hong Kong, Japan, Korea, Malaysia, Mongolia, Nepal, Singapore, Taiwan, Thailand and Vietnam.

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). Their recent efforts are on constructing large and multi-speaker (1,000 speaker/corpus) speech corpora of different speech type, region, accents, dialects and planning of platform exchange. 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. In alphabetic order, the other countries have all provided important messages to share within the community. Japan continues to be the supporting force in the organization in both research and technology development as well as community support (for example housing the official O-OCCOSDA website at ATR to date), in addition to having successfully launched both national and international projects. SITEC, Korea has successfully constructed various kinds of speech corpora of not only Korean, but also Chinese, Japanese, English and Spanish, applied them to technology development, and is standing firmly on its own feet after the initial 5-yr period. Mongolia has been constructing speech corpora and welcomes outsiders to visit. Nepal has begun similar efforts of late and also wishes more interaction from the outside world. Hong Kong, Singapore and Taiwan have all identified code mixing in technology development. Last but not least, Vietnam has presented active researches that are both language specific and up to date.

There were also 4 motions and resolutions. Motion 1 was on future meetings and host countries. There were 2 formal proposals and presentations from Vietnam and Japan, and 2 informal ones from Mongolia (in absentia) and Nepal. The resolution states that O-COCOSDA 2007 will be held in Vietnam (Dr. Luong Chi Mai) and 2008 in Japan (Dr. Sotoshi Nakamura), while during the 2007 meeting Xinjiang University and Katmandu, Nepal will be discussed for future meetings.

Motion 2 is a proposal from ex-convener Professor Shuichi Itahashi to publish a book by the 2008 meeting in celebration of the 10^{th} anniversary of the organization. The motion was adopted unanimously; the book will be coordinated by Professor Itahashi. Tentative title of the book is Standards and Resources of Oriental Spoken Language Systems. Contents will include corpora, phonetics, phonology, prosody, assessment of synthesizers and recognizers, software tools, etc. Languages to cover include Chinese (Mainland, Hong Kong, Taiwan), Indian (Hindi, Marathi, Telugu, English), Indonesian, Japanese, Korean, Malay, Mongolian, Nepali, Thai, Vietnamese, Singaporean (Chinese, English). Estimated book size or 4 papers from 17 regions at 6 pages each is around 408 pages.

Motion 3 is on regional resource sharing and

collaboration by Dr. Riza Hamman of Indonesia (in absentia). The COCOSDA handbook proposal by convener Dafydd Gibbon and reported by Dr. Nick Campbell earlier was discussed adopted as a first step towards sharing and collaboration. Dr. Nick Campbell will be the O-COCOSDA coordinator/editor of the COCOSDA Handbook. The working title is Handbook of International Multimodal Speech Resources. Topics will include: (1.) resources for regional and local languages in the domain of Oriental COCOSDA, African, Oceania, Europe, and the Americans; (2.) multilingual resources; (3.) multimodal resources; and (4.) tools for resource creation, management and processing.

Motion 4 is on membership drive, funds and newsletter by Dr. Shyam Agrawal of India. The motion was adopted and will coordinated by Dr. Agrawal. Resolutions include the following: (1.) Dr. Agrawal will contact Pakistan, Sri Lanka, Bangladesh and Bhutan on our behalf, Dr. Luong Chi Mai will contact Laos, Cambodia, and Chiu-yu Tseng will contact Thailand to solicit more participation. Volunteers are needed to the Philippines and Myanmar. It was also decided that we will still focus on East Asia to South Asia for the time being, and will not include Middle East. (2.) We will establish a membership drive that includes both individual and institution members and at the same time raise some funds for maintenance and operation. The first step will also aim for newsletters to be posted on our website. (3.) A Membership Committee is set up to carry out the above. Dr. Agrawal is the head and coordinator.

6. International Activities

O-COCOSDA was also approached by the organizers of ISCSLP 2006 (International Symposium on Chinese Spoken Language Processing, Dec. 13-18, 2006, Singapore) to bring more diversity to the largely Chinese working community. The proposal was to organize a special session on multilingual corpus development. Dr. Chiu-yu Tseng was responsible for its organization. To support the event, the ISCSLP2006 organizers donated USD2,000 to O-COCOSDA 2006, and offered 50% reduction of registration fees to O-COCOSDA participants who also attended ISCSLP 2006.

The session attracted more submission than expected and extended to include 2 full sessions of a total of 12 papers. Former O-COCOSDA convener Professor S. Itahashi was invited to the conference. Members from O-COCOSDA contributed 5 papers: 3 from India, 1 from Japan and 1 from Vietnam. The papers were very well received and generated a lot of discussion by the ISCSLP community; the O-COCOSDA community was also impressed by the state-of-the-art progress presented. More interactions also occurred outside and after the session. Both India and Vietnam expressed wishes to participate future meetings of ISCSLP. We will continue 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.

7. Future Plans

Speech research has become popular gradually in Oriental countries including Vietnam, Xinjang Uygur Autonomous Region of China, Nepal, etc. We intend to hold the Oriental COCOSDA meetings in these places in order to promote speech research there. We have also learned about the discrepancies between development of speech technology and the overall Internet infrastructure in some countries in the region. Most of the countries in the region have not yet constructed the general rails and roads of fast and stable Internet communication while researches at the labs thrive. We have also reached the time that more concrete cross-country projects on resource sharing and exchange may not be too far away. One thing for sure, we have attracted willing and eager participation from the beginning, and the community has grown steadily.

8. Conclusion

This paper summarized the development of Oriental COCOSDA and its recent activities. Continuous speech, spontaneous speech, expressive speech, speech affects, code-mixing are identified as the major topics by the O-COCOSDA community. Continued efforts on the above topics will still be worked in association with tones, pitch accents, speech prosody as well as other distinct segmental phonetic features exhibited in Indian languages. Non-alphabetic writing systems will also continue to be another research focus. These language specific features unique to the region and the ever growing necessity of speech corpora and

speech related research appear to be the drive force of O-COCOSDA.

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

References

[1] Nick Campbell, "COCOSDA – a Progress Report," Proc. LREC 2000, Athens, Greece, pp. 73-76 (2000).

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

[3] S. Itahashi, C. Tseng and S. Nakamura, "Oriental COCOSDA: Past, Present and Future," Invited paper, Proc. *LREC 2006 (The 5th International Conference on Language Resources and Evaluation*, (May 22-28, 2006), Genoa, Italy, (2006)