COCOSDA - a Progress Report

Nick Campbell

(on behalf of the COCOSDA CCC)

ATR Spoken Language Translation Research Laboratories
Kyoto, Japan
nick@slt.atr.co.jp

Abstract

This paper presents a review of the activities of COCOSDA, the International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques for Speech Input/Output. COCOSDA has a history of innovative actions which spawn national and regional consortia for the co-operative development of speech corpora and for the promotion of research in related topics. COCOSDA has recently undergone a change of organisation in order to meet the developing needs of the speech- and language-processing technologies and this paper summarises those changes.

1. Introduction

With the goals of collaborative work and information interchange for resources and standards in spoken language engineering, COCOSDA was established to encourage and promote international interaction and co-operation in the foundation areas of spoken language processing. Collaboration which transcends national boundaries is important both because of the practical and scientific value attached to systematic work which encompasses a range of languages and analytic approaches and also because of the practical need to establish common methods of performance description and quantitative comparison.

COCOSDA [1] covers all aspects of Spoken Language Resources (production, annotation, distribution, standards, etc), and of Spoken Language Systems Evaluation (speech recognition, speech synthesis, speech-translation, oral dialog, speaker recognition, language identification, etc). From its first meeting in 1990, as a satellite event of ICSLP, and arising from ideas generated at the 1989 ESCA Noordwijkerhout workshop on Speech Input/Output Assessment and Speech Databases, COCOSDA has provided a forum for international action and discussion, and gives platforms for groups of workers to exchange information and to set up collaborations in the field of spoken language engineering. Many of the world's leading workers are amongst its members and the group discussions are unconstrained by any special interests. Previous meetings have taken place in Chiavari 1991, Banff 1992, Berlin 1993, Yokohama 1994, Madrid 1995, Philadelphia 1996, Rhodos 1997, Sydney 1998 and Budapest 1999.

In 1993, COCOSDA fostered Euro-COCOSDA which, in collaboration with Elsnet for the Natural-Language and Terminology aspects, led to the foundation of the `European Linguistic Resources Association' (ELRA) [2], the European counterpart to the American `Linguistic Data Consortium' (LDC) [3]. This resulted in the first International Conference on Language Resources and Evaluation, in Granada in May-June 1998, which attracted 520 participants. We now see the formation of an Oriental COCOSDA, which has held meetings in Japan in 1998, and Taiwan in 1999, and will host the main COCOSDA meeting in Beijing later this year. The latest Oriental COCOSDA meeting included participants from Korea, China (including Hong Kong), Taiwan, Japan, Thailand, Europe, and America, and was attended by more than 125 researchers from industry and academia.

The last full COCOSDA meeting took place in Budapest, as a satellite meeting of Eurospeech-99, and papers were presented on issues regarding Spoken Language Resources and Spoken Language Systems Evaluation, including project and program reports and paper proposals on the methodological, technological, and scientific aspects related to the fields covered by COCOSDA with special focus on the theme of `Co-operative corpora: tools and materials for large-speech-database collection, annotation, and use'.

2. COCOSDA Issues

COCOSDA is concerned with issues such as the design, construction, and use of Spoken Language Resources (SLR) and of Spoken Language Technologies (SLT), both monolingual and multilingual, including the production, annotation, validation, distribution, and formatting, etc., of spoken corpora. It is also concerned with the legal aspects of the collection and distribution of speech-based resources, the application of SLR, and the analysis of user needs (both for research and industry), the definition of standards, and the provision of information regarding newly available SLRs. The annual meetings feature reports on the resources and standardisation of national and regionally sponsored projects.

Although not an official standards body, COCOSDA is active in the co-ordination and standardisation of

assessment techniques as a precursor to the development of International Standards. It encourages quantitative, comparative, qualitative and perceptive evaluation, development of measures, protocols and metrics for the situated evaluation of applications. It covers issues in SLT evaluation such as the benchmarking of systems and products, evaluation in SLT systems (speech recognition and understanding, voice dictation, oral dialog, speech synthesis, speech coding, speaker and language recognition, etc.) including systems incorporating a speech component, such as multi-modal and multimedia systems.

Multi-linguality and dialectical variation are of particular importance to COCOSDA, since in many cases the same tools and techniques can be used for different language regions.

2.1 Regional Consortia

COCOSDA is not a funded organisation; it is supported by active and concerned members of the speech and language processing communities who have an interest in fostering the development of speech and text corpora for international use. However, from these activities national and regional consortia have been formed, such as the American LDC and European ELRA, which are non-profit commercial organisations that charge for access to corpora, much like a software data publisher, producing annual CD-rom disk sets for public distribution. Institutions subscribe to these organisations, as they might do for e.g., census data, and the disks can be used by the subscribing institutions at their discretion, as with library access. Since the consortia are engaged in physical distribution of data, the usual copyright restrictions apply. The consortia pay the costs of production, using the institution's label design, and put the item in their catalogue at a price determined by the data provider, usually charging only for production costs. The copyright (if any) remains with the provider.

2.2 Special Interest Groups

In addition to spawning such commercial organisations, COCOSDA has also assisted in the formation of a Special Interest Group for Speech Synthesis. The COCOSDA Working Group on Speech Synthesis maintained an archive of references and web-sites and encouraged the collection of large single-speaker corpora suitable for research and use in prosodic analysis and speech synthesis systems.

In November '98, COCOSDA, in conjunction with ESCA, organised a four-day Evaluation and Research Workshop which was focussed on the assessment of speech synthesis systems. Participants at the workshop were invited both as listeners and as providers of TTS systems, and more than 40 systems were evaluated in parallel using common texts and listening environments. Attendance at this workshop was 50% over-subscribed (120 participants attended while only 80 were expected) and the resulting profits were donated to ESCA for the use of SynSig [4], a self-supporting Special Interest Group that was formed as a result of discussions held at the workshop.

SynSig will take over the COCOSDA speech synthesis web site, enhance the exchange of news on recent research developments and make available relevant resources (databases, corpora, tools, reference lists, etc.).. The SIG has a mailing list and home page (synsig@isca.org (for the committee) synsig@itl.atr.co.jp (direct to the members), http://www.itl.atr.co.jp/cocosda/synthesis/synsig.html) to stimulate further evaluations that benefit the science and help both researchers and business users of synthesis to improve systems to meet their needs.

Being independently funded, the SynSig will be able to allocate money for specific targets, such as student travel, web-site maintenance, disk storage, archiving, etc. It will also encourage the setup and design of co-operative international and multilingual experiments, and will organise the exchange of students and the collection and exchange of tools and resources for teaching, evaluation, and research purposes.

3. Oriental COCOSDA

The first International Workshop on East Asian Language Resources and Evaluation was held in Tsukuba, Japan during May '98, as the First Workshop of Oriental COCOSDA. The purpose of this workshop was to exchange ideas, share information and discuss regional issues; and to promote speech research on oriental languages regarding the creation, utilisation, and dissemination of spoken language corpora as well as the assessment methods of speech recognition and synthesis systems. Among the 28 contributed papers, 17 were from Japan, four from China, four from Korea and three from Taiwan. There were 54 participants during the two-day meeting. The English-language edition of the Journal of the Acoustical Society of Japan produced a special edition containing papers from the workshop [5].

The 2nd Oriental COCOSDA Workshop was held in Taiwan in May '99. It was primarily attended by academics (105) with a few representatives from industry (21), and of these most were engineers, speech scientists, linguists or phoneticians. The papers presented covered recognition, synthesis, labelling, evaluation, and system design, and the 15-minute presentations allowed for little more than advertisements for the various topics. However, the main discussion of the meeting was devoted to co-ordination issues: national, regional, and international initiatives and programmes, and to defining the new co-operative trends within language resources and evaluation. An apt parallel was drawn between `language engineering' and `aircraft engineering' by a guest speaker who pointed out that there are many more manufacturers of aircraft than there are of aircraft engines. By effectively adapting existing resources to meet the local needs, progress can be made more quickly.

In contrast to the presentations typically made at international conferences, where the emphasis is on

successful results, there was more focussed discussion of `needs', with emphasis on what CAN'T yet be done, to `learn first what needs to be learnt' in order to bootstrap corpus development from successful existing projects. In spite of language and dialect differences, the various countries in the South East Asian regions have similar task requirements, i.e., for modular processing that can be shared by different sites (e.g., for segmentation and labelling of corpora), and automatic detection of segmentation-errors).

Taking the SAMPA and ToBI labelling systems as examples of common working representations, there was agreement to standardise interfaces and to encourage the interchange and common development of models, modules, and systems (an example of this was an offer by Philips to make public its source-code for phonetic label-mapping). A satellite meeting was held on the last evening to resolve differences in the various machine-readable transcription systems for Mandarin Chinese and to prepare reports on individual features in order to unify the labelling of speech corpora and enable interchange of transcriptions.

During a panel session, discussion focussed on finding `the Right Model for East Asia', and comparisons were made with both the LDC and ELRA (the former being DARPA sponsored, and the latter being funded with EU support). It was agreed that no similar funding structures exist for the East Asian countries and that rather than relying on such top-down support, a bottom-up approach might be more effective. Initial actions will include the sharing of regular reports on national projects and experiences and the setting-up of area-based mailing lists and web-based facilities, though in some partner countries access to the internet is still difficult and permission to put government-owned intellectual property on a publicly-accessible web-site can be hard to obtain.

Oriental COCOSDA will hold regular symposia on odd-numbered years (LREC's symposia are held on even numbered years), with the goals of a) matching the quality of industrial corpora for the research domain, b) defining the needs for collaboration among East-Asian countries, and c) sharing resources, such as corpora and software, and tasks, such as assessments and projects. A by-product may be to encourage international conferences to devote more sessions to East-Asian requirements.

4. The Budapest Meeting

The 1999 COCOSDA Workshop took place on Friday, September 10th in a building close to the Eurospeech International conference centre, on the day following the main conference. About 70 people participated in the workshop. The morning was devoted to COCOSDA presentations and the afternoon meeting held jointly with members of COST-258 for discussion of "Speech Segmentation Techniques".

The COCOSDA-99 Budapest Workshop presentations included "Semi-automatic labelling of a speech database for phonetic research in the Welsh language", "User Interface of the Spoken Language Corpus CLAP", "Designing and Collection of Australian-Korean AV database", "The Transcriber Tool", "GSK: Linguistic Resources Sharing Organisation and New Projects on Speech Research/Corpora in Japan "The Spoken Dutch Corpus Project", "Effective knowledge distribution and management", "TransEdit - a transcription tool from transcribers for transcribers", and "Oriental COCOSDA - updates from East Asia". Abstracts are available from the COCOSDA web pages.

A working meeting of the COCOSDA Central Co-ordinating Committee was held in Budapest, and several changes to the previous organisation were proposed. They have since been circulated to all past and present members of the Central Co-ordinating Committee and have now been formally approved.

First, In order to avoid redundancy with regular scientific sessions on Spoken Language Resources, or on Assessment, which take place in International conferences, such as ICSLP or Eurospeech, it was proposed to establish "Topic Domains" in place of "Working Groups", and to have a "Topic Rapporteur" representing and reporting the main developments in each topic area. Among the topics suggested were "Corpus annotation tools", "Emotional speech corpora", "Multi-modal corpora", "Dialog evaluation" and "Minority languages".

The Rapporteur will have the task of gathering, between two COCOSDA workshops (that is during the period of each year), all relevant information on that topic. He or she will then have the task of organising a session at the COCOSDA workshop to report on that topic, including a presentation by him(or her)self of the survey made during the year (including if necessary a selection of what was presented at the main conference itself), and one or two presentations, either invited or after submission on that topic.

Second, regional area activities (Asia, Australia, EU, US, etc.) will be reported in the same way, with selection of rapporteurs for the different topics to be the responsibility of each CCC member. This change will allow better rotation of people and of responsibilities, and will also enable immediate action whenever there are new developments.

COCOSDA will henceforth sponsor an annual "Best Student Paper" prize for papers submitted at international conferences. The topics for the award will include assessment, or use of assessment methodologies. The prize will include a two-year student subscription to the ISCA Speech Communication Journal.

4.1 Renewal of the CCC:

Since the usual term for COCOSDA Convenorship is two years, and the present convenor has been in place for four years already, a new convenor was elected and a new post of Deputy-Convenor was created. The Deputy-Convenor will be required to convene the CCC in cases when the Convenor is unable to do so.

Professor Lin Shan Lee (National University, Taiwan) was approved as new Cocosda Convenor, and Dr Khalid Choukri (CEO ELRA, and Director ELDA) was approved as COCOSDA Deputy Convenor. Their first act was to renew the Central Co-ordinating Committee and to create a new COCOSDA Advisory body.

The goal of the latter is to ensure a good coverage of the global regions, and of the various specialist fields addressed by COCOSDA (including Corpus and SLR, Speech Input Evaluation, and Speech Output Evaluation), and to keep close relationship with former members to benefit from their experience. The following categories of member were proposed: i) Data Centre representatives, ii) Topic Experts, iii) Former CCC members and iv) other experts:

The present Advisory Committee is to consist of the following former members of the CCC:

- USA	D. Pallett, M. Liberman
- Europe	J. Mariani, A. Fourcin
- Asia	H. Fujisaki, A. Kurematsu
- Australia	B. Millar, P. Dermody
- Synthesis	L. Pols
- Recognition	D. Pallett
- Corpora	C. Castagneri

Three new CCC positions are established as below:

- K. Choukri	(Deputy Convenor)
- N. Campbell	(Permanent Secretary)
- S. Itahashi	(Rep. Oriental COCOSDA).

COCOSDA will continue to focus on the topic domains already established, and will now start to develop regional programs, such as "Regional Program-Europe", "Regional Program-Asia", etc., and will identify rapporteurs for each different region. With respect to further topic domains, in addition to the current "Evaluation of Speech Understanding/Dialogue Systems" (W. Minker), and "Multi-Modal Corpora" (S. Nakamura), "Corpus Annotation and Tools", "East Asian Languages", "Minority Languages" or "Lowly Represented Languages" have been proposed as candidates.

4.2 Forthcoming Workshop in Beijing:

The next COCOSDA workshop will take place on Saturday, October 21, 2000, as part of Interspeech (formerly ICSLP) in Beijing. Topic suggestions and information proposals can be submitted via the COCOSDA web pages.

5. Conclusion

COCOSDA can perhaps best be regarded as a pre-profit organisation for the co-operative standardisation, development, and testing of large-scale international or inter-regional projects. As an information-based organisation, it does not exist to dictate methods and directions but to provide for and to inform those who do. It can best do this by setting up working groups in the relevant areas and by encouraging the regular exchange of reports, resources, and ideas.

To facilitate such international progress, there will be increasing provision of internet-based facilities, in order to a) enable modular development and use of distributed or shared tools and resources, b) make corpora and systems available for reference and testing, and c) educate and inform through the provision of materials and the suggestion of standards.

Information is now a *commodity*, tools a *service*, and software a *product*, but by encouraging distributed and co-operative standardisation of tools and resources in the pre-competitive stages of research and development, and by fostering regional consortia when the technologies are mature, we can encourage a component-openness (of modules and interfaces, but not necessarily of source-code) and may facilitate the rapid development of speech-based systems by reassuring the owners of information that they stand to gain more than they lose by such co-ordinated actions.

References

(1) COCOSDA: www.itl.atr.co.jp/cocosda
(2) ELRA: www.icp.grenet.fr/ELRA/home.html
(3) LDC: www.ldc.upenn.edu
(4) ISCA SynSIG: www.itl.atr.co.jp/cocosda/synsig
(5) Journal of the Acoustical Society of Japan (English Edition), vol.20, no.3, May 1999.

Written records have been produced from the earlier COCOSDA Workshops, and copies of the booklets are available on request from Professor Hiroya Fujisaki of the Science University of Tokyo.

Proceedings from the Noordwijkerhout Workshop can be obtained, subject to availablility, from ISCA.

Proceedings from the Oriental COCOSDA Workshops can be obtained from Professor Itahashi of the University of Tsukuba Japan.

We would like to thank the ATR Interpreting Telecommunications Research Laboratories in Japan for the use of their internet facilities.

Note: Parts of this paper were presented in an earlier version at ICSP-99 in Korea.