COCOSDA & Oriental COCOSDA: a Progress Report

Nick Campbell

ATR Interpreting Telecommunications Research Laboratories
Kyoto 619-09, JAPAN
nick@itl.atr.co.jp, www.itl.atr.co.jp/cocosda

Abstract

This paper presents a review of the activities of COCOSDA, the International Committee for the Coordination and Standardisation of Speech Databases and Assesment 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. Recently, Oriental COCOSDA has been formed to determine the special needs of East-Asian languages and to explore the speech-related aspects of text-processing for non-roman alphabets. The paper concludes with a report of the latest Oriental-COCOSDA meeting in Taiwan.

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 cooperation 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 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, and Sydney 1998.

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), the European counterpart to the American 'Linguistic Data Consortium' (LDC). 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 parent COCOSDA meeting in Beijing next 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 next full COCOSDA meeting will take place in Budapest as a satellite meeting of Eurospeech-99, on Friday September 10, 1999, and papers are being invited on all 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. A special interest will be devoted this year to 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 assesment 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 multimodal and multimedia systems.

We foresee the future inclusion of a wide range of related technologies that are not directly speech-based, including information-retrieval (IR), automatic summarisation (AS), machine-translation (MT), internet speech (SMIL), and so forth. These will include topic detection and tracking (TDT) research, especially in cross-language mode, and TIDES, which is expected to cover as many as 30 languages. These latter technologies all have a substantial speech component (e.g., of the current 11 information sources in TDT, 8 are radio or TV, with textual input being from speech-recogniser output in the most important test conditions). This will reflect the current shift in funding towards research on multilinguality and towards applications that combine traditionally distinct technologies (such as SR and IR), or define new technologies that cross the traditional boundaries. This shift has important implications for issues of interest to COCOSDA.

Multilinguality and dialectical variation are also of particular interest to COCOSDA, since in many cases the same tools and techniques can be used for different language regions.

3 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.

In defence of the need for charges, Mark Liberman of the LDC replied on the Linguist mailing list in 1995 that *"The LDC membership fee for a university is $2,000, and for this fee everyone at that university can get an unlimited and perpetual research license for everything that the LDC publishes during the year of membership [...] We feel that $2,000, which is roughly the cost of a moderately configured PC or an international conference trip, is not out of line even for university researchers. Speaking for myself, I have a great deal of sympathy for the effort to provide research resources free or at minimal cost, [...] These efforts rely heavily on volunteer labor and other donated resources; in several cases they have also relied on cash donations from the LDC. However, volunteer labor is rarely available in the needed quantities; and of course LDC-supplied cash, as well as the existence of the LDC as an organization, depends on income from somewhere. [...] Whether a particular database is worth a certain price is a matter of individual taste, but as a matter of simple arithmetic, the fees charged are unlikely to cover all the costs incurred."*

For institutions interested in CD-ROM publication of a language-related database that is considered to be of general interest, the LDC offers to pay the costs of production, using the institution's label design, and will put the item in their catalogue at whatever price the institution chooses, charging only for production costs. The copyright (if any) remains with the institution.

Like the LDC, ELRA has been entrusted with a mission to ensure that Language Resources needed by Language Engineering researchers and developers are made available when they already exist or to produce them in

a cost-effective frame. This mission is tuned from time to time to anticipate future requirements. The following is from ELRA (with thanks to Khalid Choukri):

Such a mission can be itemized as:
- The identification of useful resources.
- The distribution activities and Pricing policy.
- The validation and Quality assessment.
- Handling the legal issues of Language Resources.
- Information dissemination, Promotion and Awareness.
- Commissioning of LR & Market watch.

In the speech area, we see that the catalogue has grown from the 22 initial resources of March '96 to more than a hundred today. However, many key resources are still not available for a large number of languages (including Western European ones), and even among Western European languages a large number of languages are still not covered. We consider our basic resources to be:

- Articulatory databases (e.g. ACCORD)
- Pronunciation lexicons (BDLEX)
- Name pronunciation lexicons (ONOMASTICA)
- Newspaper read text (BREF, Siemens-100, Apasci)
- Basic telephone speech (SPEECHDAT)
- Telephone-based speaker verification. (PolyVar)
- Text corpora for language models (MLCC, Le Monde)
- Basic speech data with some phonetic material and some phonetic sequences, by a small number of speakers, recorded in a quiet environment (EUROM 1 & BABEL)

Other concerns will be also addressed, in particular quality and validation issues for a large number of types of resources such as speech databases over the fixed & mobile telephone networks, speech databases from broadcast news, speech databases recorded in car environments, etc. A number of European initiatives (within the R&D framework programmes of the European commission) as well as a number of national programmes have already been launched to fill this gap.

4 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, 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.). Its main web site will have regional mirror-sites to facilitate accessibility. The SIG will create a mailing list (synsig@esca-speech.org) with the goal of stimulating further evaluations that benefit the science and help both individual and business consumers of synthesis to select and design systems that 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 organize the exchange of students and the collection and exchange of tools and resources for teaching, evaluation, and research purposes.

5 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, utilization, and dissemination of spoken language corpora as well as the assessment methods of speech recognition/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.

The 2nd Oriental COCOSDA Workshop was held in Taiwan in May '99. It was primarily attended by academics (105) with 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 coordination issues: national, regional, and international initiatives & programmes, and to defining the new co-operative trends within language resources and evaluation. An interesting parallel was drawn between language engineering and aircraft engines; when it was pointed out that there are many more makers of aircraft than there are of their engines.

In contrast to the presentations typically made at international conferences, where the emphasis is on successful results, there was more discusssion of 'needs', with the emphasis focussing on what *CAN'T* yet be done. There was an expressed need to 'learn first what needs to be learnt' and so to bootstrap corpus development from successful existing projects. The different coutries in the region have similar task requirements – language & dialect variability in the data, modular processing that can be shared by different sites (e.g., segmentation & labelling of corpora), and automatic detection of segmentation-errors.

Taking the SAMPA and ToBI labelling systems as examples of common representations, there was agreement on the need for standardised interfaces to encourage the interchange and common development of models, modules, and systems. One example of this was an offer by Philips to make public its source-code for 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 the 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 approaqch might be more effective. Initial actions will include the sharing of regular reports on tional projects and experiences and the setting-up of -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-acccessible web-site can be hard to obtain.

Oriental COCOSDA wll 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.

6 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, 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, which 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 provision of materials and 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 cosortia when the technologies are mature, we can encourage a component openness (of server, but not necessarily source) 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