A NATIONAL CLUSTER OF SPOKEN LANGUAGE DATABASES FOR AUSTRALIA

J. B. Millar [*], P.Dermody [•], J.M.Harrington [†], J.Vonwiller [‡]

[*] Computer Sciences Laboratory, Australian National University
[•] National Acoustic Laboratories, Chatswood, Sydney
[†] Speech, Hearing and Language Research Centre, Macquarie University
[‡] School of Electrical Engineering, University of Sydney

ABSTRACT - This paper addresses the issue of the nature of a viable national resource of spoken language in Australia. The importance of such a resource for the development of speech technology in Australia is explored against the background of the economic, political, and legal issues that have frustrated previous attempts to develop such a resource. A proposed solution is provided in the form of a cluster of technically compatible databases in which each component of the cluster will have its independently determined content arising from the primary purpose behind its collection. The primary compatibility will be that each component corpus will have the same technical structure and the same standards of data description. Secondary compatibility will arise by making the components of the cluster available under well-defined conditions to the speech technology community via a set of database nodes, each of which will be accessible by electronic data links.

INTRODUCTION

This paper is a companion paper to Millar et al (1990) in which the definition of an adequate description of spoken language data for a generally available national database is addressed. In this paper the same authors apply the general principles articulated in their previous paper in the Australian context. How can the Australian speech science and technology community move towards adopting a community-wide principled approach to the accumulation of a data resource that is also relevant to the needs of individual projects? Conformity to a rigid set of standards has a poor track-record in all manner of human endeavours, yet cooperative progress has no future without agreement on certain standards. The previous paper has proposed that a necessary first step is agreement on a standard set of descriptors of spoken language data. Once the necessary structure for describing the data is in place then data standards themselves can be clearly articulated. Rigidity is therefore minimised by replacing absolute standards by clear documentation of the quality of the data so that any shortcomings may be allowed for by the user of the data.

This paper first of all examines the role of spoken language data within the speech science and technology (SST) community, then looks at the background to acquiring such data in Australia and overseas. The concept of a national cluster of databases is then mapped onto the data needs of the Australian SST community and the proposed standards of data description are reviewed. The paper concludes by addressing the issue of the integration of existing data into the national system and the modus operandi of management of the national system.

THE ROLE OF SPOKEN LANGUAGE DATA

Spoken language data has two major roles in the SST community. First, it has a fundamental role for the speech scientist being the domain in which the scientific method is applied in order to structure our knowledge about speech on the basis of adequate and true observations of the real world. Second, it has an important secondary role for the speech technologist as the means for formal evaluation of the performance of the technology. Our scientific theories and the technology we build on them are only as good as the data on which they are constructed and the data against which we test them. Consequently a principled approach to the structuring and description of data is of utmost importance in the field of speech and language technology.

This critical dual role of spoken language data and the fact that the creation of well-structured databanks is an onerous task, implies that high standards should be set at the outset. However it must also be acknowledged that there are a large variety of uses to which collected speech data may be put, not all requiring the same standards of quality. We therefore propose mandatory high standards of description of the data but simply encourage the use of the highest standard of data that is appropriate to the application. This approach is designed to maintain maximum cooperation between data collectors in sharing their data, and to provide the user of the data with sufficient information to judge whether or not the data is of sufficient quality for their needs.

OVERSEAS PROJECTS

The last five years have seen the emergence of a number of projects that are specifically devoted to the collection of a speech database. The most advanced of these, called TIMIT, has been developed in the United States under the Speech and Natural Language Program of the 'Defense Advanced Projects Research Agency' (DARPA). This is a database of some 6300 utterances amounting to 5 hours of annotated speech. More recently, a further speech database project has been launched by the same speech technology group to collect speech of over 1000 hours duration. A project called SCRIBE, which is closely related to the DARPA project, but which is designed more specifically for the major accents of British English, has recently been started in the UK under the initiative of the SALT (Speech and Language Technology) Club. The aim of this project is to collect sentences and conversational speech from a large number of male and female speakers. The database will be made widely available to the whole of the UK speech and language community on CD-ROM in the format proposed by the Speech Assessment Methodology (SAM) project of the ESPRIT programme.

Although the English language database developments are well ahead of those of other languages, there have nevertheless also been significant advances in Europe. In France, for example, the GRECO project has been constructing a large database of spoken French (BDSONS) as well as a database of the phonological and lexical aspects of French (BDLEX). The SAM project has released a CD-ROM comprising spoken digits and some sentences in five European languages.

AUSTRALIAN BACKGROUND

The first wave of awareness of the need for a substantial spoken language resource in Australia occurred at the beginning of the current decade when two laboratory reports independently proposed or described the collection of substantial bodies of speech data from the Australian speaker community (O'Kane et al, 1982; Clark and Fraser, 1982). The scope of each of these projects has recently been reviewed by Millar (1989). One of these projects received funding for one year only, and the other received no funding at all. Funding bodies were not willing to expend resources on developing a data bank which was not a part of a clearly defined technological project which has good prospects of lucrative application. Promise for the future, general benefit to a wide-range of projects, and least of all, academic value and the pursuit of excellence in our knowledge of our spoken language contributed little to the acquiring of necessary financial support. The consequence of this attitude is that the speech research community has a body of data resembling a patch-work quilt. The lack of common methods of collecting, storing, describing and organising data is a strong inhibitor towards cooperative use of this expensive resource.

This problem with the funding of technology infrastructure is not unique to Australia. The recently reported denial of funding for the SCRIBE national speech database in the UK which is supported by a wide section of the vibrant speech science and technology there, emphasises that a monolithic approach to spoken language database aquisition for its own sake, and that of a wide-range of current and future projects, is very difficult to 'sell' in the present political and economic climate. The situation in the USA indicates that the might of powerful government, academic, and industrial committment to speech technology have been combined to generate and disseminate such databases as TIMIT (Texas Instruments, MIT, NIST and DARPA collaboration).

In the absence of such a confluence of might the Speech Assessment Methodologies (SAM) project of the ESPRIT programme in Europe has a adopted an approach of collaborative development of standards that will be adhered to by the collaborators, or which will be an agreed interchange standard, from which and to which interfaces will be created to support local standards. The SAM project has some 23 collaborating partners throughout Europe and has formally expressed interest in a partnership with the Australian speech community for the mutual exchange of data and software for data manipulation.

Developments in Australia, latent since the activity of 1982, were aroused by this interest in partnership with SAM expressed in the last quarter of 1989. A request for expressions of interest in establishing a national database in Australia was published in the newsletter of ASSTA late in 1989. Various researchers expressed interest but it was immediately obvious that there existed a wide variance in the desired outcome of such an initiative. Those expressing interest were invited to meet in Sydney in early 1990 to set up some structure to define the vision for a national database and to move forward with the vision. This paper and its companion paper are the first public airing of the initiatives taken arising from several prior discussions that were reported at the meeting, and discussions at the meeting itself. The working group initially dubbed the Australian National Speech Database Initiative (ANSDI), has recently applied to ASSTA for formal recognition as an ASSTA sub-committee thus placing it under the control and encouragement of the formally constituted body representing speech science and technology professionals in Australia.

CURRENT AUSTRALIAN ASPIRATIONS

The concept of a cluster of independent but compatible databases has received broad support within the ANSDI group and among those it has consulted. This approach releases the national database project from immediate specification of the range of content which is required to satisfy national needs. What is required is that each sectional interest defines its database needs in a way that allows implementation as a component of the national database. Each component will be optimally designed for the needs of a specialist group such as the automatic speech recognition researchers, audiologists, lexicographers, speech pathologists, language educators, phoneticians, or natural language modellers. In this way specific task- determined requirements enable confident definition of content linked to specific projects (and therefore maybe funding) while preserving overall compatibility. Within the total database, material developed for one project may be readily used in whole or in part in another project. This multiple use of data is where the national database makes its strong contribution. Contributors to the national database are therefore encouraged to see their contribution as a component of their intellectual property which is being placed in the public arena. It will be used with acknowledgement, and developed and enhanced by others using the database as a "domain of interaction" (Millar et al, 1990). In this section we survey some of those sectional areas where distinct database needs are felt.

One major aspiration is a database of the 'General Australian' variety of English on a scale commensurate with TIMIT for American English, and SCRIBE for British English. Just as compatibility between the component parts of the Australian database will be important to allow overlapping usage, there are similar compelling advantages to use linguistic materials, recording conditions and annotation schemes which will be similar in many respects to those of TIMIT and SCRIBE. Such similarities will facilitate future research on specifying the acoustic, phonetic and linguistic bases of the accent differences between General American, British English and General Australian. The ability to 'map' one accent onto another is of major importance to many aspects of speech technology: in speech recognition, for example, it would mean that a system designed to recognise General Australian could be trained on a much larger, pre-existing database of British English, or possibly even General American, while in text-to-speech, the same system could be 'switched' from one accent to another. The ability to relate different accents in this way can only be accomplished by creating a database of General Australian which is comparable, in some ways, to the already existing and rapidly emerging English language databases in the US and Europe. It should be noted however that while comparable material is desirable for comparative studies, material that is distinctive of Australian English will be necessary for complete coverage of our local speech. From the phonetic point of view, this database will be indispensable to research in automatic speech recognition, speech synthesis, and articulatory, acoustic and perceptual modelling of speech.

As there are many aspects to Australian English it is likely that the whole area may best be covered by a number of components. Individual components may cover such areas as allophones of Australian

English, intonation, discourse, sentence structure, and dialogue. Further, closely related components may be developed which are essentially extensions of existing components: for example, using the same linguistic content but extending the range of speaker types. It is likely that a series of components will be developed for the shifting patterns of migrant English.

Another aspiration is for a database suitable for audiological studies. Audiological materials have been modified for use in Australia (Clark, 1981) but control of the vital 'speaker dimension' has been at best ad hoc (Millar, 1986). Analysis of what is actually presented is becoming essential to thoroughly exploit modern evaluative and prosthetic techniques (Blamey et al, 1986).

While the multiple speaker dimension of an Australian English corpus will give contemporary estimates of the range of 'normal' speech, once it has grown sufficiently in size, it will always be necessary to focus attention on speech which is in the outer regions of the 'normal' distribution. Thus the quantitative definition of the range of normal speech will enable the identification of specific corpora of speech pathologies suitable for the study of specific types of pathology and for quantitative monitoring of the progress of therapeutic treatment.

The Australian speaker community presents many challenges to the speech scientist and will do increasingly to speech technologists. It is important that the local speech professionals are able to develop and evaluate speech technology using models of natural speech that are representative of the local speaker community. It would seem particularly important for Australian SST researchers to have access to data on distinctive aspects of English as spoken in Australia such as the widely variable diphthongisation of vowels and incidence of nasality, the distinctive patterns of intonation, and the high incidence of mother-tongue accenting of variable degree from a wide range of Asian and European languages.

The composite need of the SST community for data resources, which is incompletely covered in the above, can be met by constructing compatible components of a national database cluster. This solution, which partitions the task into manageable and hopefully fundable chunks, also preserves the opportunity to relate results in sectional areas to the spoken language performance of the whole community.

STANDARDS OF DATA DESCRIPTION

The basis of compatibility between database components is a set of common standards of data description and a common organisation of that description. The proposed file structure for spoken language corpora offered for inclusion as components of the national database was outlined in figure 1 of Millar et al (1990). The structure provided for sampled data files, environment files, window analysis files, and annotation files.

The central form of machine-readable spoken language data is a sampled data file of the the acoustic signal of speech as captured by a microphone, maybe supplemented by additional contemporary signals. The primary descriptors that must be attached to the sampled data are those contained in a file header.

Environment files specify the 'recording conditions', 'collection protocol', 'speaker characteristic', and individual factors that characterise the specific recording session. Together with the sampled data file header the content of these files define the quality of the speech data.

'Window analysis' and 'annotation' files are regarded as 'value-added' data as they are derivatives of the primary data produced as the result of machine processing of blocks of data, or human perceptual processing (or other labelling techniques) which result in a symbolic description of the data at a number of levels. The headers of these files serve to describe that processes to which they have been subjected.

It remains to specify in technical detail the content of this structure. This must be done as a focussed consultative process to ensure that the highest standards are encouraged and that its is acceptable to all contributors. The standards of description should also be acceptable to the community of users of the database which may well extend outside the national boundaries given the global use of the English language. It is therefore important that equivalence with other standardisation attempts worldwide is developed.

STANDARDS OF QUALITY

While the general philosophy of the national speech database is non-prescriptive it is also necessary for a representative group coordinating ongoing activities to act as consultants to individuals or groups interested in the generation of a recorded speech corpus and to advise about basic standards of quality. To this end the ANSDI committee intends to develop a set of basic standards related to the signal processing and environmental recording conditions to act as guidelines for future use. Similarly it is envisaged that operation notes on orthographic and phonetic transcription will be available for prospective annotators of collected data. We believe that such material will encourage the adoption of basic standards that will in turn increase the likelihood that data collected at considerable effort will be used extensively and that production of high quality and well documented data will be rewarded by suitable acknowledgement. It is anticipated that these data collection guidelines will suggest a range of generally acceptable values for a number of descriptive dimensions that are encoded in the environment files described above. Significant deviations from these normal ranges, such as speech samples collected in extreme environmental conditions, can be handled in 'questionnaire' mode within the structure of the environment files.

INTEGRATION OF EXISTING DATA

The generally non-prescriptive nature of the proposal also opens the way for the integration of existing data which can be described according to currently agreed standards and manipulated into the agreed data organisation. A survey of current data holdings and the conditions under which they were collected is also proposed and funding for the survey has been granted by ASSTA.

MANAGEMENT PROPOSALS

The management of the data is critical to its widespread acceptance and use. It must be efficiently managed allowing flexible access to its highly complex structure. It is proposed to evaluate the use of the 'Oracle' DBMS at selected sites to allow interrogation of the data and the generation of output 'datasets' in form required by the user. It must also be managed securely, hence it is proposed that all data is stored on more than one site. It must be accessible to the user community, hence it is proposed to evaluate the use of AARnet as a means to interrogate a description of the database at one or more sites, and also for data retrieval in modest quantities.

It is widely accepted that the effort needed to create a high quality structured data resource is considerable. It is therefore of utmost importance that this effort is appropriately recognised. In the past a view of data as simply 'raw material on which we operate' has caused those who have expended great effort and expense in creating a significant corpus to restrict access to it. In the national database we propose to create a 'high view' of well- organised data. Legal rights to intellectual and industrial property must be upheld and respected. Users will be required to acknowledge the work of the data compilers and where appropriate contribute to its cost and/or sign a non- disclosure agreement. Contributions to cost could well be on a sliding scale depending on the level of commercial intent of the user. However it is clearly valuable to allow free access to a descriptive database within which a potential user of the main database may examine the kind of data that is available and be advised of methods and cost of accessing the speech data itself. This descriptive database may involve access to the headers of data files or maybe only to filenames in which are coded the necessary information. This can be organised within the DBMS.

The physical means of transfer will be dictated by the facilities available at the distribution site and may attract a handling charge depending on the effort involved. The physical media could include industry standard computer magnetic tape (high handling cost, medium capacity), cartridge tape (medium handling cost, high capacity), optical disc (high medium cost, low handling cost, high capacity), floppy disc or diskette (low capacity, low handling cost, low medium cost). An optimum arrangement could be the use of diskettes for the order of a megabyte of data, or a cartridge tape for up to one or two gigabytes. The use of AARnet for data transfer as well as interrogation of the database is to be investigated.

CONCLUSION

The value of a national speech database is beyond question. It is clear that in the USA and in Europe progress in many aspects of speech technology, and in particular automatic speech recognition, is largely dependent on a well-structured speech database. A carefully structured national speech database also provides a focus for collaborative research both between institutions and across disciplines. Collaborative research between the speech and the language processing communities has been hampered in the past by the absence of a database of continuous speech: in general, the speech research community has based its research on isolated sentences or words, while research in natural language processing and lexicography has tended to use written, rather than spoken, data. An unfortunate consequence is that very few creative links are forged between the sub-disciplines of spoken language study. The sort of collaborative research between speech and language processing which could be stimulated by an emerging continuous speech database might be the closer analysis of the links between discourse analysis, prosodic structure, and low-level phonetic rules. The lack of research in this area has has had unfortunate repercussions for many aspects of speech technology: for example, the higher level processing in text-to-speech systems is usually restricted to a crude form of syntax which results in an underspecified prosody and an unnatural synthetic speech quality. Beyond the applications designed to improve directly the performance of speech and language technology, a speech database will be relevant to many different areas of basic research such as the characterisation of Australian English and the study of speaker characteristics. A thorough understanding of these topics will eventually make significant contributions to future speech technology.

REFERENCES

Blamey,P.J., Dowell,R.C., Clark,G.M., Seligman,P.M. (1987) *Acoustic parameters measured by a formant-estimating speech processor for a multiple-channel cochlear implant*, Journal of Acoustical Society of America, Vol. 82, pp.38-47.

Clark,J.E. (1981) *Four PB word lists for Australian English*, Australian Journal of Audiology, Vol. 3, pp.21-31.

Clark,J.E., Fraser,H. (eds) (1982) *Australian Speech Archive*, Occasional Papers, Speech and Language Research Centre, Macquarie University.

Millar,J.B. (1986) *Quantification of speaker variability*, Proc. First Australian Conference on Speech Science and Technology, Canberra, pp.228-233.

Millar,J.B. (1989) *Design and use of a national speech database*, Proceedings of ESCA workshop on 'Speech Input/Output Assessment and Speech Databases', Noordwijkerhout, the Netherlands, 20-23 September.

Millar,J.B., Dermody,P., Harrington,J.M., Vonwiller,J. (1990) *A national database of spoken language; concept, design, and implementation*, Proc. International Conference on Spoken Language Processing (ICSLP-90), Kobe, Japan, 18-22 November.

O'Kane,M., Millar,J.B., Bryant,P. (1982) *A database of spoken Australian English: design and collection*, Technical Note No.6, School of Information Sciences, Canberra College of Advanced Education.