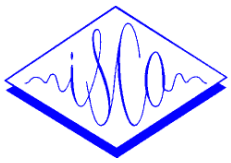


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

A NATIONAL DATABASE OF SPOKEN LANGUAGE: CONCEPT, DESIGN, AND IMPLEMENTATION

J. B. Millar, Computer Sciences Laboratory, Australian National University

P. Dermody, National Acoustic Laboratories, Chatswood, Sydney

J. M. Harrington, Speech, Hearing and Language Research Centre, Macquarie University

J. Vonwiller, School of Electrical Engineering, University of Sydney

ABSTRACT

A model is proposed for the building of a national resource of spoken language data in the form of a cluster of compatible databases. Each component of the cluster will have its own linguistic characteristics dependent on the primary purpose behind its collection. However each component corpus will have the same structure and the same standards of data description. The emphasis is on adequate description of the data rather than on conformity to a standard of recording conditions, data storage, or linguistic content. This paper outlines the rationale for such a database and proposes principles for the structuring of data storage, and for the description of important dimensions of such spoken language data. Some attention is also given to the management of such a data base within the speech and language technology community.

INTRODUCTION

The strong development of research in speech science and the application of speech technology in Australia within the last decade has prompted a number of national initiatives: First, a series of speech science and technology (SST) conferences, then the formation of the Australian Speech Science and Technology Association (ASSTA), and now the Australian National Speech Database Initiative (ANSDI). Millar (1989) described some antecedents of current spoken language database development in Australia and pointed towards a national initiative. During 1990 an ANSDI task force has been established to prepare the ground for the creation of a national database of the spoken language of Australia. In this paper the executive members of that task force propose guidelines on the necessary common infrastructure for such a cooperative national enterprise. In a companion paper (Millar et al, 1990) they will look in more detail at the nature of the databases that can be constructed and used within this structure. The need for an Australian National Database of Spoken Language, seen against the international backdrop of speech and language technology development, is reviewed only briefly here.

The present proposal for a cluster of Australian databases can be seen, in part, to be complementary to initiatives that have developed in the last 5 years specifically devoted to the collection of a speech database: TIMIT in the USA, SCRIBE in the UK, BDSONS in France, and SAM in many countries of Europe. Australian English has some 13 million speakers (Millar, 1989) who exhibit considerable phonological stability. It has three overlapping varieties broadly based on socio-economic factors, and very few regional differences. Thus its speakers form a moderately large and uniform speech community which is ready to utilise the benefits of speech technology. In addition there is a substantial population of nearly 4 million first generation migrants who speak a wide range of accented English. This speaker community presents a special challenge for Australian speech technology.

The creation of databases with similar material, recording conditions, and annotation schemes as those developed for other English language databases in the UK and USA will facilitate future research on specifying the acoustic, phonetic and linguistic bases of the accent differences between American, British and Australian varieties of English. A clear definition of these differences will enable the effective tuning of speech technology systems to operate efficiently in such varied linguistic environments.

Previous experience indicates that the creation of a single monolithic database to suit the needs of all speech and language professionals is a task of such enormous proportions that even a valiant and well-funded attempt will never fully meet the needs of all sectors of interest. An alternate approach has been adopted in which acceptance of a basic set of protocols for the description of speech data and its labelling becomes the foundation for a cluster of compatible databases. This concept releases the project from immediate specification of the range of content which is required, allows progress to be made in areas where specific task-determined requirements enable confident definition of content, while preserving the feature of compatible extendability as resources and requirements develop.

In this paper we examine the dimensions of the necessary common infrastructure and make preliminary proposals about how each of these dimensions may be structured. Most of these dimensions will be of a technical nature but it is recognised that there are political (aspects that motivate people to cooperate), legal (aspects that protect intellectual property rights), and economic (aspects that enable resources to be channelled to enhance development) dimensions of the project. These can be treated only briefly in this short paper. First of all we offer a perspective on the role of data and databases in speech and language research.

DATA AS A RESOURCE

Well-structured data is a resource that traditionally has been under-valued. Our theories and the technology we build on them are only as good as the data on which they are constructed. Consequently a principled approach to the structuring and storing of data is of utmost importance in the field of speech and language technology. As spoken language data collection is such a onerous task it is important that we set high standards 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. We therefore propose high standards of description of the data rather than high absolute standards of quality that could never be acceptable to all. This approach is designed to maintain maximum cooperation between data collectors in sharing their data, and to provide the user of the data with enough information to judge whether or not the data is of sufficient quality for their needs.

DATA AS A DOMAIN OF INTERACTION

The proposed flexible cluster of databases will have far greater potential than simply a data resource of estimable quality. Agreement on protocols for data description immediately opens up the scope of use of the database as a domain of interaction between different research groups and different disciplines. This interaction can take place in the form of "value-added" extensions of the database which build on the material collected or analysed by others. These value-added extensions could be in the form of analyses performed using sophisticated computational facilities or specialised human expertise. In this way such additions as "formant tracks", "perceived stress markers", and various levels of segmentation and labelling would be submitted in "interchange format" for inclusion as a value-added component of the existing database.

While this openness to additions may appear as potential to compromise the quality of the database, the principle of separation of analytic and annotational data into separate files, and rigorous description of the processes performed should prevent any negative impact and allow for very positive additional value when the quality of additions is high. With this structure in place it will be possible to encourage not only ad hoc growth in the richness of the database but also planned projects of cooperative enrichment of the data.

This would be an important development in Australia where groups of colleagues that are sufficiently large and diverse to apply all the relevant skills to the processing of spoken language data are widely spread geographically. Effective use of this database as a medium and as a catalyst for this cooperative approach to research and development would have a major impact on our progress in the speech and language technology field.

DATA IN CONTEXT

Spoken language data always has a context which is often responsible for subtle variations that appear in its low-level analysis. It is of utmost importance that the information regarding the context of a segment of speech is accessible by the user of the data. This proposal suggests a structure for retaining the fullest reference to context via pointers to "environment files", and to any prior temporal segmentation of the data from a given recording session.

The environmental contexts deemed important are the nature of the speaker (stopping short of identity), the nature of the speech task in which the speaker is involved, and the ambient conditions in which the speaker performed the task. In addition to the general description of these three environments there will be specific variations which pertain to a particular recording session, these also must be attached to the data. Technical proposals for storing and attaching these environmental data to the speech data are given in the next section.

In addition to the environmental context the immediate temporal context of the speech stream itself is highly significant. An important principle to apply to the design of all segmentation and excision operations in order to create new database components is that fragments of speech data should always be defined in terms of their source file and their position within that file. For example, isolated words extracted from a recording seession in which they were embedded in a carrier phrase must be stored with the filename of the digitised version of the full session plus a pair of numbers which delimit the position of the word within that file, whereas a read passage may require the original filename plus several pairs of numbers to select portions of the file avoiding hesitations and misreadings which are immediately corrected. While only the words, or only the correctly read passage may be of interest to the automatic speech recognition researcher, the tempo of the carrier phrase or the nature of spoken misreadings and correction style may be of great interest to the phonetician or auditory psycholinguist.

TECHNICAL DIMENSIONS OF DATA DESCRIPTION

The linguistic content of a speech corpus is its most obvious characteristic, and yet, apart from a few indicative comments, it will feature minimally in this paper. It is on the linguistic dimension that most component parts of the national database will differ. 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. However, while each component will be designed with the linguistic requirements of one group in mind, its content will be a window onto the speech of the Australian community. While recognising diversity at a linguistic level, it is the aim of ANSDI to stimulate cooperation and unity in all other dimensions of database design and development so that overlap at the linguistic level may be fruitfully exploited.

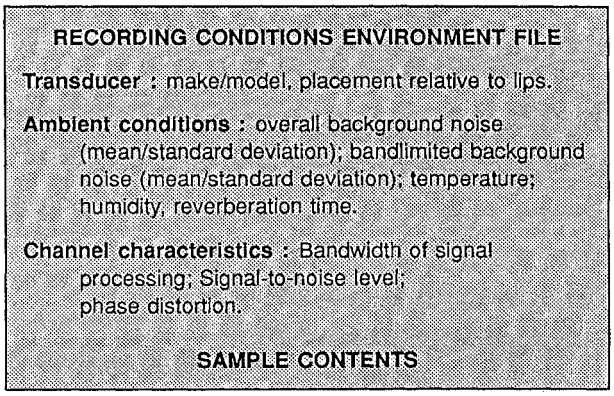
Data structure and data description

The proposed file structure for spoken language corpora offered for inclusion as components of the national database is illustrated in figure 1, in which hatched boxes represent those files which are proposed to be mandatory. All others are optional.

Sampled-data files. The basic form of spoken language data is the acoustic signal of speech as captured by a microphone. This may be supplemented by additional contemporary signals such as those captured by an electro-glottograph, aerometer, electro-palatograph, or alternative microphones under different channel conditions - each of these alternatives would have a unique entry in the file header and a unique character in the file name. These data are classed as "sampled data" being the result of a sampling process operating on continuous signals transduced from the physical world. This class can also include any transform of these signals which leaves the signal as a regularly sampled one-dimensional function of time.

The primary descriptors that must be attached to the sampled data are contained in a file header. The header information defines Sampling Rate, Quantisation level, Compression code (if used), name of the file containing the original unsegmented material and pointers to the location of the segment in the original file, plus pointers to four mandatory "environment" files.

Environment files. The first environment file will specify the "recording conditions" (RC-environment) which are all those static conditions of the environment in which the signal was produced. This will address the acoustic environment, the recording equipment used (including options selected), and the position of the speaker with respect to transducers. It may also include specifications of the anti-aliasation filter such as its bandwidth at full amplitude, attenuation at half-sampling rate, and type of anti-aliasation filter.



The collection protocol (CP-environment) file will specify the speech task that the speaker is asked to perform. This will include all relevant factors describing the way in which speech was elicited from the speaker, encompassing written and spoken instructions and prompts.

COLLECTION PROTOCOL ENVIRONMENT FILE

Elicitation technique : Sequential reading; Flash Card; Screen prompt; Spontaneous monologue; spontaneous dialogue; verbal task, competing task.

Item delimiters : New breath, Short gap; Continuous.

Type of material : word, short phrase, sentence; discourse.

Discourse Style : Procedural; Narrative; Hortatory; Expository.

SAMPLE CONTENTS

The speaker characteristic (SC-environment) file will specify the speaker in terms of their linguistic, educational, geographic, and health background, as well as age, relevant interests (music) or habits (smoking), plus any other long-term influences on their voice. In all cases the data held should not be sufficient, of itself, to identify the speaker as an individual person, therefore "contact" information must be held manually outside of the database.

The "individual session" file will record any anomalies relating to recording conditions, collection protocols, or speaker behaviour that were specific to the session in which the basic signal file was derived. This will include adjustments to equipment, variations in protocol due to errors and recovery, and variation in the speaker due to a minor illness, tension, exhaustion, etc.

SPEAKER CHARACTERISTICS ENVIRONMENT FILE

Physical : Sex, Age, Height, Weight, General health.

Linguistic : Place of birth, Age of entry to Australia, Mother tongue, Languages spoken (fluency)

Educational : Level, Institution, Place, Dates

Occupational : jobs, places.

Audiological : status, date tested.

Health history : relating to respiration, vocalisation, hearing, ear, nose, throat, & chest (e.g. surgical procedures, bronchitis, asthma,hayfever, tonsillitis, sore throat, sinusitis, chronic colds, post-nasal drip)

Influences on voice : Voice/singing training, regular high vocal effort, etc)

Musical ability : Instrument, Level

Smoking history : When, How much.

Family history : Father, Mother, Spouse (Place of birth, Age of entry to Australia, Mother tongue, Education, Occupation)

SAMPLE CONTENTS

These environment files will be structured to allow maximal description to be coded where desirable, but will have a subset of mandatory fields without which the data will have little validity beyond its original use.

Value-added files. Beyond the sampled data files and environment files there will be two "value-added" forms of data files. The first are transforms of the data that are derived by the application of algorithms which operate on a "window" excised from the sampled data. These will include a range of spectral transforms using specific models of the input signal (e.g. Fourier, Auto-regressive, Wigner-Ville, Pitch-extractors, Formant-trackers). These will typically contain packets of "N" values, where each packet is derived from a window on the basic signal which is incremented through the signal on an arbitrary time-base. The window position increment value will be held in the file header. The headers of these files will not be fixed length descriptive headers as used with the sampled data files, but rather "history" headers which contain all information to allow them to be recreated from their sampled-data origins using specified algorithms and parameters. These history headers will need to be extendable as multiple processes may be involved.

The second value-added type of files are symbolic annotation files of the basic sampled-data signal. These could range from the orthography of the utterance with end-points equal to the beginning and end of the file, through to highly detailed phonetic transcription time-locked to the signal at many points. Higher level symbolic annotation of either the basic signal file or of the orthographic or phonetic transcription of the utterance must also be provided for within this structure. These higher levels will include prosodic markers, syntactic markers, discourse structure markers and the like. An annotation will comprise a "label" (being the value of the annotation within some annotation scheme), and a "location" parameter which can indicate the temporal binding of the label to the time course of the sampled data file. More than one kind of location parameter will be needed. They may range from "start,stop" markers, to "here" markers, or even more fuzzy "here, plus or minus so-much" markers. The headers of these files will need to include reference to the annotation scheme used and the annotator responsible for applying the scheme to these data.

It should be noted that these data-structure proposals are for a standard "interchange format" suitable for storage, manipulation, and access in a large data-base management system (DBMS). It is clear that some transformation may be necessary when applying processing software which has conflicting requirements for header information or for file naming. For instance, it will be necessary to use software which will enable large amounts of acoustic data and their associated analyses to be accessed by phonetic context. Some care will be necessary to ensure that such software, or modifications thereof, is compatible with the component databases that we create. It is strongly recommended that a standard format, which all serious users are prepared to use for interchange, be established.

In addition to developing standards that are acceptable to all national contributors, they should also be acceptable to the community of users which may well extend outside the boundaries of the nation, given the global use of the English language. It is therefore important that direct equivalence with other standardisation attempts worldwide are maintained.

MANAGEMENT OF THE DATA CORPUS

There are many aspects of management of the data corpus which need to be decided and about which we can only give some general hints as to the direction that seems optimal from our current perspective.

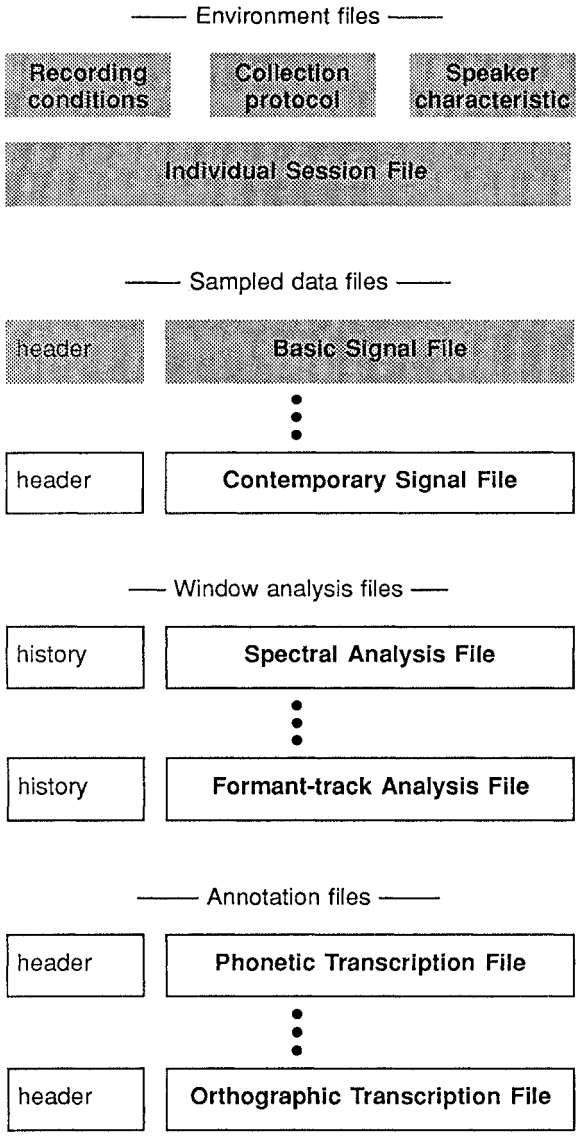


Fig.1 - Proposed file structure for spoken language data

Access, Security, Search. As a machine-readable corpus of spoken language is a highly complex data structure, powerful data management techniques will be required to allow adequate access to the data. It is therefore recognised that the data should be organised such that it can be managed by a sophisticated DBMS such as "Oracle". As only certain sites may have such facilities, the "interchange format" of the data should be ideally independent of the DBMS.

It is proposed that all data held in the national database system be held on multiple sites for security purposes, and that a full descriptive database of the contents of all components of the national cluster be held at one or more sites which can be accessed via the Australian Academic Research network (AARnet). It would seem valuable to allow free access to the 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.

Intellectual/Industrial Property. If the national database is to succeed it must address the issue of reward for the effort expended in collecting data, adding value to that data by means of further processing or annotation, and placing it in "interchange format". It is not unreasonable to expect some small fraction of the cost of creating a corpus, or of adding value to a corpus by means of further processing, to be borne by the user. This cost may be on a sliding scale depending on the level of commercial intent of the user. The actual transfer of data from the database to a user should allow for creation of an invoice, and the attachment of a notice of acknowledgement which details the protocol for acknowledgement to the work of those whose labour brought the data into being.

Data transfer. 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.

We expect to see a gradual focussing on specific signal digitisation facilities, audio-visual data verification facilities, and archiving media as the value of building on the work of others is practically realised. The building of interfaces between popular speech workstations and speech signal processing packages and the agreed "interchange format" will assist this process.

CONCLUSION

We have proposed a model for the building of a national database of spoken language which is designed around a set of principled protocols which effectively define the most important dimensions of spoken language data. The national database itself will be a cluster of component databases which are independent in content but which have a common underlying structure. This common structure is important in order to facilitate access to the data using a common software interface. This interface will allow unified search and operational activities. Search activities will primarily be developed via a DBMS, allowing reports to be generated on the existence of data according to various search-constraints. Operational activities will be developed by users who wish to transfer data to their own facilities, analyse the data, and maybe create further value-added components for submission to the database.

A major characteristic of this proposal is that it is aimed at cooperation on a national scale, and that it will transform a host of small incompatible speech corpora into a truly accessible and useful national resource.

REFERENCES

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 cluster of spoken language databases for Australia", to be presented at the Third Australian International Conference on Speech Science and Technology (SST-90), Melbourne, 27-29 November.