SPEECH KNOWLEDGE,STANDARDS AND ASSESSMENT

presented on behalf of the SAM consortium by

Adrian Fourcin and Jean-Marc Dolmazon
University College London & Institut de Communication Parlee

INTRODUCTION

This brief overview is designed to provide background information for the poster related to the work of the SAM 'Speech Assessment Methods' Project (2589) - concerned with the design and application of multi-language EC standards. At present the project is based on the collaboration of twenty-six laboratories in eight countries, six within the EC and two from EFTA. The project is now at the start of its third year in ESPRIT II. This follows a preliminary 'Definition Phase' (ESPRIT 1541) in which the status of work in the area, and the requirements in Europe and the rest of the world were investigated, and a 'bridging' 'Extension Phase', in which preparatory work for the Main Phase was undertaken.

Current work is in progress in three inter-connected working areas:

I Speech Recognition Assessment (Input)

II Speech Synthesis Assessment (Output)
III Enabling Technology and Research (ETR)

At the beginning of the SAM Project, the need to ensure a practical basis for ready collaboration between so many different laboratories in different countries was met by the definition of a reference, standard, workstation - SESAM. The minimum hardware requirements for SESAM are an IBM pc-at or compatible computer, an analogue interface board, OROS-AU21 or AU22, 1 Mbyte of extended memory, and means for accessing speech data eg CD-ROM reader. C is used as the common programming language. Each one of three workgroups, above, has made use of this simple reference standard so that software, data and assessment results can be interchanged. This has proved to be very successful both between project members and in the provision of data and support for other laboratories across Europe - all the work of the SAM Project is designed to be readily available within the European Community.

I INPUT ASSESSMENT

In recognition assessment, the simple reference standard workstation has been implemented and tested in multi-lingual, multi-laboratory trials.

In order to provide a flexible tool for recogniser assessment, the component software packages are designed as separate modules which can be independently developed by different laboratory groupings within the project. The first package, PAOSAM, is designed to be capable of managing the information associated with the standard SAM format speech databases. The second package, EURPAC, primarily controls the interaction between the assessment system and the recogniser itself and the third package. This last package, SAM_SCOR, provides a series of performance measures. All three software modules are interconnected via ASCII files, and all programs are in C using the microsoft 5.1 compiler, and executable on the SESAM workstation running MS-DOS as the operating system

Database Management

The RISE program has been developed to cater for the major needs of data retrieval and data archiving for all languages and all speakers in the SAM project. A commercially available DataBase Management System (ORACLE) is used as the basic building block. The management structure has been designed to allow the integration of all of the characteristics of both present and future SAM speech databases. Effectively, RISE enables the user to specify the characteristic assessment aspects to be targetted in terms, for example, of language, speaker and speech types, and for an automatic procedure to be utilised for the composition of training and testfiles in the assessment of a defined recogniser.

Control Module

The EURPAC program is designed to operate from this basis in controlling the assessment of isolated or connected word recognisers. The assessment session can be controlled by information given in a separate control file, defined by the user, and giving details of the unique serial number of the test run, the identification of the recogniser, and the names of the configuration-, training-, test- and response-files. An important aspect of the design of this particular software module is that it uses resident drivers to control individual recognisers. In this way, the greater part of the software is quite independent of the analogue interface board which is utilised, and it is easier to develop new recogniser drivers which can have separate communication protocols.

Scoring

The SAM_SCOR program provides a range of recognition performance measures - hit; miss; substitution; correct rejection; false alarm. In addition, at the isolated word level, confusion matrices, confidence analyses, and the application of the McNemar test are standard facilities. For connected word and continuous speech recognisers string matching at the orthographic level is available employing NIST scoring routines which have been made executable on the SESAM workstation. The output of this scoring software is designed to provide uniform presentations of the assessment results that are easy to understand and cross compare. SAM_SCOR generates a file which can subsequently be fed back into the DBMS to make it possible to relate speech material characteristics to recogniser performance measures.

Applications

More than 10 EC laboratories in the Project have been involved in the application of recogniser assessments so far for six commercially-available or in-house recognisers. Considerable use has been made of the first SAM CD-ROM speech database - EUROM 0 - which gives 5 hours from 20 speakers in five languages. This cross laboratory single and multi-language testing of equivalent recognisers has provided the foundation for the setting up of a basic calibration procedure for the SESAM input assessment workstation. Work is currently in progress to define a common method for standard reference calibration and hardware setting up protocols.

In collaboration with the ETR Group, a new multi-lingual speech database has been designed and is in the process of being recorded. The contents of the database have been defined to meet the present and near future need for the development of diagnostic and predictive assessment methodologies. The database is divided into two sets: a 'Many Speaker set' and a 'Few Speaker set'. The vocabulary of the 'Many Speaker set' contains a list of selected numbers between zero to nine thousand nine hundred and ninety nine covering all the phonotactic possibilities of the languages' number systems, and blocks of five sentences giving continuous speech with paragraph prosody rather than individual sentences. The vocabulary of the 'Few Speaker set' is expanded with a CVC list and more repetitions per item.

II OUTPUT ASSESSMENT

Standard word-level and sentence-level segmental multi-lingual intelligibility tests have already been defined. They can be automatically generated on the SESAM workstation in the languages of the project using phonotactic and word frequency constraints. Compatible software provides for response collection, collation and scoring.

Segmental Structures

The SAM segmental test contains guidelines for the automatic generation of nonsense-word lists for all eight partner languages, using a set of fixed word structures and phoneme lists. The test material is language specific in that phoneme combinations respect phonotactic constraints for the languages in which they are prepared. The SAM group has chosen to use nonsense words in its definition of this standard with an open response set in order to get an intelligibility score which is not influenced by contextual information or semantically restricted answer choices. This type of material is the most relevant when an analysis of phoneme confusions is required, and in application, for instance, to synthesis material where error patterns may be quite device-specific. The SAM Segmental test consists of two parts: a first "core test" containing structures common to all languages of the consortium, and which cover consonants in initial, medial and final positions: VCV, VC (+ fixed final V for Italian) and CV. In all cases, the full inventory of consonants is used with only a sub-set of vowels. This sub-test cannot be considered as a *full* diagnostic test, but it is substantially diagnostic for consonants. The "full test" will include more extensive language-specific and even synthesiser-specific sub-tests with complex structures such as CVC, CVVC, VCCV and possibly CnVC, CVCn. Phonemes are presented in equal numbers per list so that an equal probability score will be obtained. This score can then be weighted according to phoneme-frequency-of-occurrence counts to obtain scores which reflect phonemic balance.

Segmental Assessment

A system to support the automatic segmental assessment of synthesisers has been implemented on the SESAM workstation. The subject responds using the keyboard, results are then automatically scored, to produce percentage scores, confusion matrices, analysis in terms of certain types of phonetic feature, eg place of articulation and voicing, and the effect of vowel environment.

Assessment of words in context

A test of word intelligibility in sentence context has also been developed for SESAM, using semantically unpredictable sentences (SUS). Grammatical structures and word lists are defined for all the languages of the consortium to permit the generation of an unlimited number of test sentences. Work on prosodic assessment is also in progress.

III ENABLING TECHNOLOGY AND RESEARCH

The core SAM workstation, SESAM, has been specified and implemented for data collection, following standard protocols, database management, and speech signal labelling. A phonemic notational system for all European languages, SAMPA, has been developed and is in use both for manual labelling and, currently, for semi-automatic label alignment. Phonemic level structural constraints across the languages of the project have been compiled and are used in corpus definition. Broader descriptors are being investigated for multi-lingual application. Other, physical, levels of description are being quantified as a contribution to analytic methods of assessment. Information on cross-language lexica is being compiled.

SESAM

Hardware (see the INTRODUCTION above) and software specifications are now well established and widely applied in regard, for example, to: the structure and code normalisation of software; the formatting of data and organisation of databases; and the provision of interfaces.

Two, key, software packages are central to the use of the workstation within the project. The first is EUROPEC, which is designed to provide for the realisation of large speech databases. Two-channel acquisition (eg for microphone and laryngograph signals) and monitoring is now possible with visual prompting for the speaker which may be manually controlled or automatically triggered as a function of signal level. Automatic end-point detection facilitates the handling and recording of large organised corpora. This is also substantially assisted by the automatic inclusion in the database of description text files in standard form with header and body, so that the orthographic prompt can be routinely incorporated together with complete sessional and recording item and condition information.

A complementary package, VERIPEC, is designed to give ready access to these standard data and text files, making it possible to display the orthographic prompts, access and monitor recorded items and show the label files.

The second important package, PTS, is designed fo operate from the data acquired via EUROPEC. Its primary function is to enable the labelling of speech data files with either SAMPA (see below) or IPA notations, using window based displays of waveform and spectrograms of the signal.

Data

The first SAM database, EUROM 0, was distributed on a single CD-ROM and contained five hours of speech material recorded using a condenser microphone in anechoic rooms from four single accent speakers in each of five languages (with 16 kHz sampling). NATO single and triple digit sequences were obtained with only the speech signal, and a continuous speech passage, with a common numeric theme across languages, was recorded using two channels - with both speech and laryngographic inputs.

A new database is now in preparation using the same standard format with sixty speakers in each of eight languages. Nonsense words, number sequences up to 9999, (both phonotactically balanced) and situationally linked sentence blocks are being recorded. Anechoic condenser microphone recordings will permit the subsequent imposition of post-production effects. A small subset of data will have two channel representation, as above. Two CD-ROMs are planned for each language - using 20 kHz sampling.

SAMPA

The SAM Phonetic Alphabet (SAMPA), which defines a standard keyboard based notation (ASCII) corresponding to the relevant International Phonetic Association symbols for each of the languages represented in the project, was agreed very early in the project, and has now been extended to cover all the major European languages. It has also been adopted by a number of ESPRIT projects and both the British and German national speech databases. This consensus for the representation of phonemic contrasts in all the languages of the group provides a common labelling basis for cross-comparison and for a structured multi-lingual approach to database specification in the development of standard methods of assessment. The basic SAM transcription system was originally intended to evolve as a multi-tier labelling tool and work is currently directed towards the introduction of prosodic and acoustic element levels of description.

Labelling

Multi-lingual labelling, in which phoneme categories are assigned to successive regions of the speech signal, has always been an important part of the SAM group's activity. This is because overall assessment, detailed evaluation and the processes of training themselves ultimately depend on an accurate definition of speech which can be given in phonetic and orthographic terms. So, although the precise assignment of discrete categories, for different sound classes, to the continuous speech signal is an impossible task - since the subjective level of labelling is not compatible with any physical set of exact temporal stretches of the signal - the consistent correlation is of real value. The SESAM workstation is designed to support this work, and manual labelling in all the languages of the project has provided essential reference material.

A further development of this work currently involves a semi-automatic approach, using label alignment. In this way the larger quantities of speech material generated by current database gathering, and which are in need of labelling, can be accommodated without imposing an impossibly large manual task.

In Conclusion - The project has provided an opportunity for multi-language work in Phonetic Sciences,which would not otherwise have been possible and we are glad to stress the collaboration and goodwill which have made the work, across Europe, so effective.