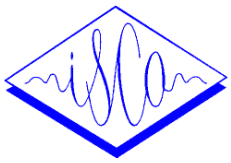

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

EUROM - A SPOKEN LANGUAGE RESOURCE FOR THE EU

The SAM Projects

Dominic Chan [1] Adrian Fourcin [1] Dafydd Gibbon [2] Bjorn Grandstrom [3] Mark Huckvale [1]
George Kokkinakis [4] Knut Kvale [5] Lori Lamel [6] Borge Lindberg [7] Asuncion Moreno [8]
Jiannis Mouropoulos [4] Franco Senia [9] Isabel Trancoso [10] Cor in 't Veld [11]
Jerome Zeiliger [12] (in alphabetical order)

[1] *Dept. of Phonetics and Linguistics, University College London, UK;* [2] *Fakultat fur Linguistik und Literaturwissenschaft, Universitat Bielefeld, Germany;* [3] *KTH, Sweden;* [4] *University of Patras;* [5] *Norwegian Telecom Research, Norway;* [6] *LIMSI, France;* [7] *Center for PersonKommunikation, Denmark;* [8] *Univerisitat Politechnica, Catalunya;* [9] *CSELT, Italy;* [10] *Inesc, Portugal;* [11] *Speech Processing Expertise Center, The Netherlands;* [12] *Institut de la Communication parlée, France*

ABSTRACT

A summary of the progress of development and the current realisation of a CDrom based spoken language resource for 11 languages of the European Union is given; the physical conditions basic to its acquisition are defined and the criteria guiding its poly-language structures briefly outlined.

1. INTRODUCTION

Work on the planning of a poly-language resource for the Spoken Language Engineering (SLE) needs of the European Union was first started in a concerted fashion in 1986, in the preparation phase of an ESPRIT project (SAM - Speech Assessment Methods). At the very beginning of collaborative work within the SAM project, the basic aims defined were concerned with the definition of common European approaches to system assessment and corpora acquisition.

Two main lines of activity have resulted from the collaborative work directed towards the establishment of a comprehensive spoken language resource. The first line of work has defined and given rise to the implementation of a family of spoken language workstations - SESAM. The second line of work has been associated with the progressive development of databases which have been called EUROM to signify the availability of European language material on read only memory disks provided by CD-roms. In each of these two lines of activity complementary objectives have structured the way that the work has proceeded;

- The need to obtain cross comparable testing and training by the use of common material
- The definition of comparable structures of phonemically and syntactically similar material which could be realistically associated with similar levels of difficulty in use
- The definition of levels of material complexity ranging from the segmental through to the word and the sentence levels which were coherently organised across languages
- The use of common themes in connected discourse
- The definition of ranges of speakers and speaker types in a fashion which enabled cross comparison between different language resources to be achieved
- The definition of methods of phonemic/phonetic transcription which would be computer compatible and linguistically appropriate and acceptable for the purposes of accurate work in each language and for the purpose of accurate cross comparison across languages
- Agreements in regard to the use of common method of coding, storage and dissemination of data
- Agreement and common activity in respect of the definition and application of working standards

Underlying aims of this first substantial concerted European activity towards common work on spoken language resources were: to provide contributions towards developments in spoken language engineering; to focus activity in respect of a description of European languages which was directly linked to data; and to give a foundation for basic work in the speech sciences.

In the definition phase of the SAM project these considerations led to a CDrom-based five language corpus (Danish, Dutch, English, French, Italian), EUROM_0 [2] which was distributed in 1987/8. Digits and passages were recorded constituting 52 minutes of material for each language and all the material was labelled using the SAMPA transcription protocol [3]. Laryngograph signals were also recorded simultaneously with the speech pressure signal.

This earlier work has now led to the production of an eleven language corpus EUROM_1 which includes Danish, Dutch, English, French, German, Italian, Norwegian and Swedish coming from within the main phase of SAM [5] described in a little detail below and Greek, Portuguese and Spanish within SAM-A (the successor to SAM).

The present status of the EUROM_1 database is that CDrom masters are complete for Italian and nearly complete for each of the other eleven languages and funding is organised for the production of 200 pressings of each. The first set of eight languages will have their data production supported from a

VALUE Programme initiative (coordinated by UCL); whilst the remaining three languages are supported by funding from ELSNET/RELATOR (coordinated from LIMSI). The VALUE support is designed to lead to the availability of this SAM format data and an associated SESAM workstation.

2. THE DESIGN OF EUROM_1

2.1. Speech materials

Four main types of corpus material have been used [4, 6]:

1. C(C)VC(V) material in isolation and in context
2. 100 selected numbers from 0-9999, such that the main phonotactic possibilities of the language number system were covered
3. 40 short passages each containing 5 thematically connected sentences with themes common to all languages
4. sentences where necessary composed to compensate for the phoneme-frequency imbalance in the passages
5. 5 pairs of context words for use with C(C)VC(V) material

2.2. EUROM_1 corpora

For each language, different sections and differing amounts of this material were recorded and structured into three target corpora subsets [4]:

1. Many Talker corpus (MANY) - (30 women, 30 men): 100 numbers, 3 passages, 5 sentences
2. Few Talker corpus (FEW) - (5 women and 5 men selected from MANY): 5 x C(C)VC(V) material, 5 x 100 numbers, 15 passages and 25 sentences
3. Very Few Talker corpus (VERYFEW) - (1 woman and 1 man selected from FEW): C(C)VC(V) material embedded in 5 different context phrases and 5 x context words

In all languages, FEW and VERYFEW recordings have been made using both laryngographic as well as condenser microphone sensors.

2.3. Speaker selection

The speakers were selected with the aim of there being an equal number of women and men, as good a coverage of age groups, and as wide a range of normal voice types as possible. It was recommended that one main phonetic group per country should be selected in the first instance together with a small number of speakers from other accent regions [1].

3. THE RECORDING OF EUROM_1

3.1. Protocols and file formats

A detailed description of the recording protocols and file formats used is given in [3, 6].In brief, speech pressure and laryngograph signals are digitized at a sampling frequency of 20KHz and 16 bit resolution and put directly onto hard-disk. The signals are also recorded onto DAT tapes at the same time for backup; recorded data are in SAM format and the SAMPA transcription protocol is used for phonemic annotation.

The recording software EUROPEC(V4.0) produces a description file in which the recording materials are labelled at prompt level (*i.e.* a marker at each prompting instance).

3.2. Recording equipment

The recordings were made using the same type of equipment in each centre in an anechoic room environment:

- A B&K half inch microphone
- A B&K digital sound level meter
- An OROS AU21/AU22 A/D board
- A DAT recorder
- A Laryngograph
- Calibrators for the rooms, microphones and laryngographs
- A standard SESAM workstation was used to ensure that common protocols were employed

3.3. Calibration signals

The following calibration signals were recorded through the recording channels for each main data set:

- A B&K 1KHz 4230 calibrator signal through the recording microphone
- A rectangular wave of 20Hz with mark-space ratio of 4:7 through the recording microphone circuit and also via an artificial neck through the electrode circuits of the laryngograph apparatus
- silence with the microphone pre-amplifer input terminated by a 50 Ohm impedance.
- silence with the microphone in its normal position
- standard balloon burst impulse signal inside the anechoic chamber

4. THE REALIZATION OF THE EUROM_1 DATABASE ON CD

4.1. Introduction

The disciplined acquisition of speech data for the EUROM_1 corpora proved to be a major undertaking which was not without its problems of which the greatest was a six months funding delay following the project start date. A substantial degree of consistency was, however, achieved both in regard to the physical conditions of recording and in respect of their spoken language structures. Comparable phonetic structures across languages were used for CVC (and CVCV) type material. The sentences and passages were based on identical themes based on original English texts freely translated into each of the operant languages. All the recordings were based on read texts which are included together with the data files in both orthographic and in SAMPA phonotypical forms.

VALUE Programme support has separately made it possible to produce 5 CDroms (x200 pressings) for each of the first group of eight languages in order to support the use of the EUROM_1 data and the promulgation of the associated SESAM workstation. In order to make the best use of the available material it has been necessary in some cases to use a lossless

data compression program *SHORTEN* [7]. The following brief summaries outline the main components of each of the separate language based recording sets.

4.2. Danish EUROM_1 database

MANY corpus (60 speakers: 35 men, 25 women)
- 4 passages
- 5 sentences
- 100 numbers

FEW corpus (10 speakers: 5 men, 5 women)
- 15 passages
- 25 sentences
- 100 numbers x5
- 114 CVC words x5

VERYFEW corpus (2 speakers: 1 man, 1 woman)
- 114 CVC words embedded in 5 different carrier phrases
- 10 carrier words x5

The speakers were primarily recruited from university staff, students and their relatives. There are in total 25 women and 35 men speakers.

There are no laryngograph signal files in the CD distribution of this database.

4.3. Dutch EUROM_1 database

MANY corpus (64 speakers: 30 men, 34 women)
- 3 passages
- 5 sentences
- 100 numbers

FEW corpus (10 speakers: 5 men, 5 women)
- 15 passages
- 25 sentences
- 100 numbers x5
- 66 CVC words x5

VERYFEW corpus (4 speakers: 2 men, 2 women)
- 66 CVC words in context
- 10 context words x5

The speakers were recruited from TNO/IZF staff, with some PTT Research employees. These were complemented with 12 women students who were selected in order to even the balance between men and femake speakers. In total, there are 30 men and 34 women speakers.

Laryngograph signals and their associated speech pressure signals were both extracted from DAT tapes in digital form, compressed and additionally put into the last CD of this database.

4.4. English EUROM_1 database

MANY corpus (60 speakers: 30 men, 30 women)
- 3 passages
- 5 sentences
- 100 numbers

FEW corpus (10 speakers: 5 men, 5 women)
- 15 passages
- 25 sentences
- 100 numbers x5
- Isolated C(C)VCs

VERYFEW corpus (2 speakers: 1 man, 1 woman)
- contextualised C(C)VCs
- 10 context words x5

The speakers were primarily recruited from university staff, students at UCL and staff at NPL. There are in total 30 men and 30 women speakers.

All laryngograph signal files are present in the CD distribution of this database.

4.5. French EUROM_1 database

MANY corpus (60 speakers: 30 men, 30 women)
- 3 passages
- 5 sentences
- 100 numbers

FEW corpus (10 speakers: 5 men, 5 women)
- 10 passages
- 25 sentences
- 100 numbers x5
- CVC x5

VERYFEW corpus (4 speakers: 2 men, 2 women)
- CVC in context
- context words x10

In total there are 30 women and 30 men speakers in this database.

All laryngograph signal files were separately compressed and put into the last CD of this database.

4.6. German EUROM_1 database

MANY corpus (63 speakers: 33 men, 30 women)
- 5 passages
- 100 numbers

FEW corpus (10 speakers: 6 men, 6 women)
- 20 passages
- 100 numbers x5
- CVC x5

VERYFEW corpus (2 speakers: 1 man, 1 woman)
- CVC within context words
- 10 context words x5

The speakers were primarily recruited from university staff and their social contacts. In total, there are 30 women and 33 men speakers. The passages provide sufficient sound distribution and additional sentences were not needed.

Part of the passage recordings of the MANY corpus were phonemically labelled.

The CVC laryngograph signal files of the FEW corpus are not in the CD distribution of this database.

4.7. Italian EUROM_1 database

MANY corpus (60 speakers: 30 men, 30 women)
- 1 calibration passage
- 3 passages
- 5 sentences
- 100 numbers

FEW corpus (10 speakers: 5 men, 5 women)
- 15 passages
- 25 sentences
- 100 numbers x5
- 93 CVCV

VERYFEW corpus (4 speakers: 2 men, 2 women)
- 93 CVCV words embedded in 5 different carrier phrases
- 9 context words x5

In total there are 30 women and 30 men speakers in the database.

The only laryngograph signal files contained in the CD distribution of this database are those for the carrier phrases.

4.8. Norwegian EUROM_1 database

MANY corpus (60 speakers: 30 men, 30 women)

- 3 passages
- 5 sentences
- 100 numbers

FEW corpus (10 speakers: 5 men, 5 women)

- 15 passages
- 25 sentences
- 100 numbers x5
- CVC words x5

VERYFEW corpus (4 speakers: 2 men, 2 women)

- CVC words within 5 context
- 10 context words x5

The speakers were recruited so that at least the dialects in the four biggest cities in Norway: Oslo, Bergen, Trondheim and Stavanger were present. There are in total 30 women and 30 men speakers.

All laryngograph signal files are present in the CD distribution of this database.

4.9. Swedish EUROM_1 database

MANY corpus (60 speakers: 30 men, 30 women)

- 4 passages
- 5 sentences
- 100 numbers

FEW corpus (10 speakers: 5 men, 5 women)

- 15 passages
- 25 sentences
- 100 numbers x5
- 82 CVC words x5

VERYFEW corpus (4 speakers: 2 men, 2 women)

- 82 CVC words embedded in 5 different carrier phrases
- 10 carrier phrase words x5

Speakers with standard middle Swedish Stockholm dialect were recruited. There are in total 30 women and 30 men speakers.

There are no laryngograph signal files in the CD distribution of this database.

5. SESAM SLE WORKSTATION

The definition and use of common standards and tools for system assessment and data acquisition and management were core aims of the SAM collaboration. Because they were to be applied in many differing environments, a conceptual "meta" platform for their notional implementation, SESAM, was also defined. Initially, the SESAM PC based workstation was merely one practical implementation of this common approach but it proved to be of such great value in achieving consistency in joint tasks that the present CDrom work has been based on the use of the EUROPEC acquisition tools using the PC SESAM facility. These are now also being implemented for common use in the WINDOWS environment by ICP as a first step towards the next generation of SESAM facilities. The range of SESAM algorithms and SAM protocols is accessible via UCL WWW [6].

6. FINAL DISCUSSION

The systematic introduction of standard protocols for data acquisition, formatting and management which are part of the SAM de facto standards are applicable to quite new developments. The current advent of the EAGLES standards group and the inception of the European Language Resource Association already complements this earlier work and provides the means for its fuller development. Current database realizations require very large speaker populations, for example those associated with the Polyphone and SpeechDat exercises. The global, largely non-analytic, character of these databases reflects, however, the short term nature of the work itself and quite different strategies must be adopted for the longer term. Although the poly and multi-language European communication environment poses greater difficulties than those obtaining in more mono-lingual communities, the same basic problems in respect of dialect, accent and individual speaker differences remain to be solved. Further broad development is essential in order to get more viable spoken language man-machine interaction. This leads to the need for a progressive evolution of corpora types and methods of annotation, which will make it feasible to address levels of human communication; giving in turn, systematic knowledge of the automatic processes of adaptation and inference which underlie the ordinary listener's ability to tune in to previously unheard speaker-dialect and accent combinations.

This collaboration between workers in eleven European countries, with centralised help and hindrance, has made it possible to make a first substantive step, towards these ends, by the provision both of an important spoken language resource and the framework for its systematic use and extension in the future.

For further information on EUROM, SAM software and SAMPA, please e-mail to *eurom@phon.ucl.ac.uk* or address UCL WWW [6].

REFERENCES

[1] W.J. Barry and A. J. Fourcin. Esprit Project 2589 SAM-UCL-001: Speaker-selection criteria. University College London, Jan. 1990.

[2] SAM Consortium. Esprit Project 1541 Extension Phase, Final Report. University College London, Feb. 1989.

[3] SAM Consortium. Esprit project 2589 SAM-UCL-018: Speech acquisition and annotation protocols and index of mnemonics. University College London, Nov. 1992.

[4] SAM Consortium. Esprit project 2589 SAM-UCL-G004: Final report. UCL, Mar. 1992.

[5] SAM & SAM-A Consortia. VALUE Program Final Report, CCM 429: European Language CD rom SpeechDatabase Workstation. University College London, Dec. 1994.

[6] http:\\www.phon.ucl.ac.uk

[7] Tony Robinson. SHORTEN: Simple lossless and near-lossless waveform compression. CUED/F-INFENG/TR.156, Cambridge University Engineering Department, UK, 1994.