General Technical Information

In this file technical information is given on how to use the wave forms files present on the CDROM. File format together with file naming in use in the EUROM1 speech database are also described. General information about the content of the EUROM1 database is provided in the EUROM1.TXT file and information about the content of each CD in the CDnDOC_F.TXT file as well (with n running from 1 to 5 (CD1à CD5)).

General information about the CD-ROMs content

The EUROM1 database files are in accordance with the European ESPRIT N°2589-SAM recommendation, the main characteristics of which are listed below:

- digitised speech samples are gathered in a simple binary file.
- every speech sample file is associated an ASCII description file; this association relying on the file names.

File naming conventions

The name of the associated description file is identical to the name of the signal file it is associated to, except the third letter of the extension where respectively the 'S' character stands for Signal and the 'O' character stands for Orthographic.

- SAM filename example:

BFN41329.NFS <- 'S' character stating the file is a signal type one speaker
language of the corpus (F for French)
code type of the items uttered (N for Numbers)
corpus code
recording file number

The first character in the filename extension indicates the type of the items that are uttered:

D	isolated Digits,
L	Letters,
N	Numbers,
S	Sentences,
P	Passage,
W	isolated Words.

The second character in the filename extension indicates the language of the items that are uttered:

D	Danish
E	English
F	French
G	German
H	Dutch
I	Italian
N	Norwegian
S	Swedish

The third character in the filename extension indicates the type of the file itself, presently here 'S' for signal file. File types defined in the SAM recommendation, use the following characters:

S	speech wave forms file
L	laryngographic signal file
T	orthographic prompted text,
A	time-aligned acoustic events file
O	associated description file
B	time-aligned Broad phonetic labelling file
N	time-aligned Narrow phonetic labelling file
P	time-aligned Prosodic labelling file

The serial recording number of the file, consisting of 4 digits (from 0001 to 9999), guarantees the uniqueness of the filename all through a recording campaign.

- example: the filename of the associated description file corresponding to the previous signal file:

BFN51650.NFO <- 'O' character stating the file is an associated description one

Associated description file format

All the description files associated to a signal file adopt a unique format which has been defined for the labelling files. Each line is made of a specific mnemonic followed by the corresponding value. The structure consists of a header and one (or several) label bodies:

The label header from LHD: .... to LBD:

the label body from LBD: .... to ELF: (end of label file)
or a new label body LBD: ....

Here is an example of the structure:

```
LHD: keyword for the header start
FIL:        :
TYP:        :
 :          :
 :    contents of the header: information on the recording session
 :          :
 :          :
LBD: keyword for the label body start
LBR:        .
LBR:        :
 :     orthographic label body: items are labelled during the
       recording process; each line starts with the mnemonic LBR: or
       EXT: when it stands for an extension line (rest of the item)
 :          .
LBR:        :
LBR:        :
LBD: keyword for the start of a second label body
LBB:        .
LBB:        :
 :     broad phonemic label body, each line starts with the mnemonic
       LBB: or EXT: if it is an extension
 :          .
LBB:        :
LBB:        :
ELF: keyword for the end of label file
```

Content of the various fields for the different types of labels

LBA: for acoustic events
- four fields : start, centre, end, symbol.

LBB: for broad phonemic label
- four fields : start, centre, end, symbol.

LBC: for a comment in the labels
- one field: text of the comment.

LBE: for extra-linguistic event
- for fields: start, centre, end, specification.

LBN: for narrow phonetic label
- four fields : start, centre, end, symbol.

LBO: for orthographic label
- four fields : start, centre, end, text.

LBP: for prosodic label
- four fields : start, centre, end, category.

LBR: for orthographic label, at the recording process

- six fields: start, end, input gain, minimum level (negative) maximum level, orthographic description.

Summary of keywords and fields

LHD:	header keyword + version
FIL:	file type
TYP:	specific file type
DBN:	database name
VOL:	database volume ID
DIR:	directory (for the source file)
SRC:	source file name
CMT:	comment
TXF:	name of the text file
CMT:	comment
SAM:	sampling rate
BEG:	labelled sequence start position
END:	labelled sequence end position
RED:	recording date
RET:	recording time
REP:	recording place
SNB:	number of (8-bit) bytes per sample
SBF:	sample byte order
SSB:	number of significant bits per sample
RCC:	recording conditions code (reference to a recording conditions file)
NCH:	number of channels
SPI:	speaker information: sex, age, native language
PCF:	protocol file name (recording protocol used)
CMT:	comment
EXP:	labelling expert
SYS:	labelling system
DAT:	date of completion of labelling
SPA:	SAMPA version
CMT:	comment
LBD:	label body keyword
LBR:	label type containing sequence beginning (in sample), sequence end, input gain on recording, minimum sample value, maximum sample value, orthographic text prompt.
EXT:	line extension
LB2:	label type containing sequence beginning (in sample), sequence end, input gain on recording, minimum sample value, maximum sample value, corresponding to the 2nd channel (option)
LBL:	idem, when the second channel holds laryngographic signal (option)
DSC:	indicates a discontinuity in the signal file.
ELF:	end of label file keyword

Integration of the data in a database management system

In order to make easier the integration of these data into a DBMS, through automatic loading procedures, some original files have been placed on disk

number 1, in the directory EUROM1_F\DOC\SOURCES.
These files are the source files as used by the recording software (EUROPEC) and are therefore more easily computable.

The relevant files are:

SPEAKERS.DBF	file for the speakers
CORPUS.DBF	file for the corpus
*.TXT	corpus text files
CORPUS.PHO	phonetic transcription