DELIVERABLE IDENTIFICATION

Identification number	LE2-4001 - SD1.3.3
Type	Technical Report
Title	Validation criteria
Status	Draft
Deliverable	SD1.3.3
Work Package	WP 1
Task	Task 1.3
Period covered	T01-T06
Date	05/11/1997
Version	1.9
Number of pages	25
Author	Henk van den Heuvel, SPEX
Workpackage (WP)/ Task (T) responsible	WP 1 - Richard Winski, Vocalis / Task 1.3 - Kamran Kordi, GEC
Project contact point:	Harald Hoege Siemens AG, ZFE T SN 5, D-81730 München phone: + 49 89 636 3374 fax: + 49 89 636 49802 e-mail: hoege@habicht.zfe.siemens.de
CEC project officer	Mr. J. Soler
Status	Public
Actual distribution	Consortium and CEC
Supplementary notes	

Key words	Telephone, speech, database, validation
Abstract	The objective of this document is to make explicit the criteria that SpeechDat(II) databases should fulfil. The document gives an overview of the database features that are checked and of the criteria employed to accept or reject a database. Specific criteria for acceptation or rejection of a database that were developed during discussions at several workshops (Athens, Aalborg) are included in this document and can only be found here. Other relevant criteria, dealing with format specifications, transcription & lexicon conventions, and speaker coverage have already been described in other reports and are only summarised here in order to present a complete overview of the validation plan.
Status of the abstract	Public

Received on	
Recipient's catalogue number	

DOCUMENT EVOLUTION

Version	**Date**	**Status**	**Notes**
1.0	11/12/96	First draft	Sent to all partners for comments
1.1	23/12/96	Draft	Minor modifications were included
1.2	24/01/97	Draft	Some modifications added
1.3	21/02/97	Pre-final	Updated after SD1.3.1 review
1.4	28/02/97	Review	Updated after SD1.3.1 review (Brugge version)
1.5	04/03/97	Pre-final	Updated after last SD1.3.1 review
1.6	05/03/97	Pre-final	Missing speaker & environmental criteria added
1.7	01/04/97		Last updates after Review meeting Brugge
1.8	23/05/97		Table 6 updated.
1.9	05/11/97	Final	Last corrections/clarifications of label mnemos

CONTENTS

Comment:

0. INTRODUCTION

The objective of this document is to make explicit the criteria that SpeechDat(II) databases should fulfil. The document gives an overview of the database features that are checked and of the criteria employed to accept or reject a database.

These criteria were not invented by the validation centre, SPEX, but they evolved from a long line of experiences with other databases and discussions amongst SpeechDat partners. The validation criteria developed within the SpeechDat(M) project were used as a starting point. They were evaluated in a previous report (SD1.3.0, [6]) . In that report the results of the validation of (most of) the SpeechDat(M) databases were summarised and recommendations for the validation of SpeechDat(II) databases were given. These recommendations were discussed during a workshop in October 1996 in Aalborg, Denmark and a few modifications were made during a workshop in March 1997 in Brugge, Belgium. The decisions made there are included in the present document.

Apart from very specific validation criteria (like the number of permissible missing files) the databases should also fulfil a lot of other requirements that immediately follow from the specifications of the databases described in other deliverables in this Work Package. These specifications are related to the database format and structure, to the transcription conventions, speaker demographics, environmental conditions, and the lexicon contents. A summary of these specifications is contained in the present document, as their fulfilment is of immediate importance for the acceptability of a database. However, this is only a summary of these specifications and details will have to be looked up in the deliverables concerned, the most relevant of which are SD1.3.1. on format specifications [7], and SD1.3.2 on the transcription and lexicon conventions [8]. Other relevant reports are those on database contents [1,2,3], and on the environmental and speaker-specific coverage of the databases [2,4,5]. It is important to note that the specifications in these deliverables prevail over the summary in the present report in case of contradictions (due to version mismatches or other circumstances).

In succession we address the validation criteria for the following topics:

1. DOCUMENTATION
2. DATABASE STRUCTURE , FILE NAMES, AND CONTENTS
3. DATABASE ITEMS AND COMPLETENESS
4. ACOUSTIC QUALITY OF THE SPEECH FILES
5. ANNOTATION FILES
6. LEXICON
7. SPEAKER INFORMATION AND DISTRIBUTION
8. RECORDING CONDITIONS
9. TRANSCRIPTION

Where relevant a subdivision in validation criteria for fixed network databases (FDB), mobile network databases (MDB), and speaker verification databases (SDB) will be made in the above mentioned sections.

The Swiss French FDB and SDB were recorded well before the database specifications were completed. Their design may, therefore, deviate from that of the other databases. The 500 speaker databases will also not be checked strictly against the criteria in this report.

1. DOCUMENTATION

The DESIGN.DOC, in English, includes the following information:

- contact person: name, address, affiliation;
- the number of CDs;
- the contents of each CD;
- the layout of the CD-ROMs;
- formats of the speech files and of the label files
- file nomenclature and directory structure
- a specification of the individual items of the prompting material, including the selection of items to be prompted;
- specification (and motivation) for the sheet design (e.g. how items were spread over the sheet to prevent list effects);
- sheet example;
- recording platform and telephone link description;
- speaker demographics information:
  - which regions, how many of each;
  - a reasoned description of the regional pronunciation variants that are distinguished;
  - which age groups, how many of each;
    sexes: males, females (also children), how many of each;
- recording conditions: environments, networks, handsets
- annotation information:
  - procedure used;
  - quality assurance;
  - a list of non-standard and alternative spellings (or reference to file SPELLALT.DOC);
  - non-standard character set used for transcription (which must be one of the ISO-8859 sets);
  - any other language-dependent information such as abbreviations, proper name conventions, contractions (July or july, isn't, cannot or can not, etc.);
  - annotations symbols for non-speech acoustic events other than the standard defined (i.e. [fil], [spk], [sta], [int]);
  - list of non-standard symbols used to denote word interruptions and break-offs;
- lexicon information:
  - procedures to obtain phonemic forms from orthographic input (obviously not the set of rules used, because they are lab copyrighted),
  - analysis of frequency of occurrence of the sub-word units represented in the phonetically rich sentences (either phones, diphones, triphones) - a phoneme distribution is obligatory,
  - any other language-dependent information or conventions
- indication of how many of the files were double checked by the producer together with percentage of detected errors;
- any other information useful to characterise the database.

A template file with section headers and directives of information to be put into each (sub)section was distributed among partners.

2. DATABASE STRUCTURE, FILE NAMES, AND CONTENTS

2.1 Directory names and file names

The databases should all obey the following three-levels directory structure:

\<database>\< block>\<session>

Where:

<database>	Defined as: <name><#><language code>. Where: <name> can be FIXED, MOBIL, VERIF for fixed, mobile and speaker verification databases <#> is 0 for SpeechDat(M) and 1 for SpeechDat <language code> is the ISO 639 2-letters language code
<block >	Defined as: BLOCK<nn> where <nn> is a progressive number from 00 to 99. Block numbers are unique ones in all CD-ROMs and there will be as many blocks as are needed to fill a CD-ROM. These numbers must be the first 2 digits used in <nnxx> described below.
<session >	Defined as: SES<nnxx> Where <nnxx> is a progressive number from 0000 to 9999, that is the numeric call identification number also encoded in each filename.

Table 1: Directory structure

In any case both signal files and label files have to be put in the terminal node subdirectories.

In addition to the previous structures the following directories will be used to store the other (non-speech data) files:

\ *(root)*	the README.TXT file reporting the first description of the database, the DISK.ID file and the COPYRIGH.TXT file
\<database>\DOC	documentation
\<database>\TABLE	speaker, session, recording condition and lexicon tables
\<database>\INDEX	index files, e.g. contents file, corpus contents files, corpus list files, ...
\<database>\PROMPT	prompt sheet if present (with appropriate sub-directory structure if needed);
\<database>\SOURCE	any source code supplied

Table 2: Directory structure for non-speech data files

All these support files have to be duplicated on each CD-ROM.

The filenames should correspond to the following template:

DD NNXX CC. LL F

where:

DD	Database identification code (00-ZZ) For SpeechDat: A1=fixed net, B1=mobile, C1=speaker verification
NNXX	Recording session progressive number (0000-9999)
CC	Corpus code (A0-Z9) obtained by collating the corpus and the item identifiers
LL	Two letters ISO 639 language code (e.g. IT=Italian, DE=German, ...)
F	File type code O=Orthographic label file, A=A-law speech file

Table 3: SpeechDat file name conventions

NNXX in filenames may not be in conflict with BLOCK and SES numbers in path name.

Empty (i.e. zero-length) files are not permitted.
For each label file there must be one speech file and vice versa.

The correct item codes should be used (see tables 5 and 6):

All lines in text files should end with <CR><LF>. This holds for all label files, all table files, all index files (including CONTENTS.LST), and the SUMMARY.TXT file.

All table files and index files should report the field names collected in each record as the first row (header) of the file. In this header the same tab spacings should be used as in the rest of the file.

2.2 The DOC directory

The following files should be in \<database_name>\DOC:
. DESIGN.DOC
. TRANSCRIP.DOC (optional)
. SPELLALT.DOC (optional)
. SAMPALEX.PS
. ISO88591.PS or ISO88592.PS or ISO88597.PS
. SUMMARY.TXT
. SAMPSTAT.TXT

The validation of the DESIGN.DOC main documentation file is described in section 1. TRANSCRIP.DOC (if present) should contain the transcription instructions to the transcribers. SPELLALT.DOC (if present) should contain a list of alternative spellings used for specific words. ISO88591.PS, ISO88592.PS or ISO88597.PS is a postscript file containing the ISO-8859 character table used for transcription. SAMPALEX.PS is also a postscript file; it lists the SAMPA symbols used for the phonemic transcriptions in the lexicon. SAMPSTAT.TXT is the output of the acoustical check on the speech files performed by each partner.

The SUMMARY.TXT ASCII file should describe all recorded material in the database.

The file should have the structure below:

1. the full directory in which the speech and label files are to be found;
2. the session number;
3. a string of typically N codes, where N is the number of total items collected in the database. If an item is present, its 2-char corpus code should appear in the corresponding position; if the item is missing, a '--' should appear. Of course a database with more than the agreed N items should have all codes present. The N codes must simply be concatenated with no intervening characters (such as spaces) in between.
4. the recording date
5. the recording time of the first item
6. optional comment text if desired

All fields are separated by **spaces**.
An example of a SUMMARY.TXT file can be found in SD1.3.1, section 8.2 [7].

An additional file, VALREP.TXT, containing the validation report will be created and added by the validation centre.

2.3 The TABLE directory

Tables should be in \<database_name>\TABLE
. LEXICON.TBL
. SPEAKER.TBL or SESSION.TBL (both for SDB)
. REC_COND.TBL (obligatory for SDB only)

The validation of LEXICON.TBL is dealt with in section 6; the validation criteria for the SPEAKER.TBL and SESSION.TBL files are given in section 7; and the validation of REC_COND.TBL is described in section 8.

2.4 Other directories

Index files (optional) should be in \<database_name>\INDEX. Only CONTENTS.LST is mandatory.

The index files (if present) obey the nomenclature
<database><language_code><item_code>.LST, e.g. A1ENN3.LST.

CONTENTS.LST file contains the following information:

- CD-ROM volume name (VOL:)
- full path name (DIR:)
- speech file name (SRC:)
- corpus code (CCD:)

- corpus repetition (CRP:) [1]
- speaker code (SCD:)
- speaker sex (SEX:)
- speaker age (AGE:)
- accent of caller (ACC:)
- orthographic transcription of the uttered item (LBO:)

The fields are delimited by [TAB]s.
It is directly derived from the label files by reading the given mnemonics. The software tool used to generate CONTENTS.LST can be included in the CD-ROMs to allow users to create customised index files. An example of a CONTENTS.LST file can be found in SD1.3.1, section 7.1 [7].

Prompt sheet files (optional) should be in \<database_name>\PROMPT.

Delivered program code should be stored in \<database_name>\SOURCE.

The next section supplies a table summarising all files required, copied from SD1.3.1.

[1] With respect the original specifications on the SpeechDat(M) deliverable D.1.4.1 “Specification of Telephone Speech Data Collection”, the fields “CCD:” and “CRP:” have been added. The first one is allows a better accessibility to the data and the second one is needed for SDB collections.

2.5 Summary of files required

File(s)	Sect.	FDB	MDB	SDB
speech files	4	Yes	Yes	Yes
label files	5	Yes	Yes	Yes
session information file	6.3	one of these	one of these	Yes
speaker information file	6.1	or both	or both	Yes
recording condition file	6.2	-	-	Yes
pronunciation lexicon file	6.4	Yes	Yes	Yes
contents index file	7.1	Yes	Yes	Yes
corpus contents index files	7.2	-	-	-
speaker list files	7.3	-	-	Yes
readme file	8.1	Yes	Yes	Yes
disk id file	8.1	Yes	Yes	Yes
copyright file	8.1	Yes	Yes	Yes
summary file	8.2	Yes	Yes	Yes
design documentation	8.2	Yes	Yes	Yes
transcription manual	8.2	-	-	-
ISO-8859 table	8.2	Yes	Yes	Yes
SAMPA table	8.2	Yes	Yes	Yes
aternate spelling list	8.2	-	-	-
source files	8.2	-	-	-
prompt files	8.2	-	-	-

Table 4: Overview of mandatory files per database type. Sect. refers to the related section numbers in SD1.3.1.

3. DATABASE ITEMS AND COMPLETENESS

3.1 Mandatory items

The mandatory database items for FDB and MDB databases are listed in tables 5 and 6. Spontaneous items are shaded. These tables were copied from SD1.3.1.

Corpus identifier	Item identifier		Corpus contents	
A	1-3	1-6	3 application words (6 for MDB and FDB collections with a small number of speakers - corpus codes from A1 to A6)	
B	1		1 sequence of 10 isolated digits	
C	1		1 sheet number (5+ digits)	4 connected digits
C	2		1 telephone number (9-11 digits)	
C	3		1 credit card number (14-16 digits)	
C	4		1 PIN code (6 digits) (set of 150 SDB codes)	
D	1		1 spontaneous date, e.g. birthday	3 dates
D	2		1 prompted date, word style	
D	3		1 relative and general date exp.	
E	1		1 word spotting phrase using an application word (embedded)	
I	1		1 isolated digit (2 for MDB with codes I1 - I2)	
L	1		1 spontaneous, e.g. own forename	3 spelled word (letter sequences)
L	2		1 spelling of direct. city name	
L	3		1 real/artificial for coverage	
M	1		1 currency money amount	
N	1		1 natural number	
O	1		1 spontaneous, e.g. own forename	
O	2		1 city of birth / growing up (spont)	
O	3	3-4	1 most frequent cities (set of 500) (2 for 250 speakers MDB with codes O3 and O4 - set of 25)	
O	5	5-6	1 most frequent company/agency (set of 500) (2 for 250 speakers MDB with codes O5 and O6 - set of 25)	
O	7		1 “forename surname” (set of 150 SDB “full” names)	
Q	1		1 predominantly “*yes*” question	2 questions, including “*fuzzy*” yes/no
Q	2		1 predominantly “*no*” question	
S	1-9		9 phonetically rich sentences	
T	1		1 time of day (spontaneous)	2 time phrases
T	2		1 time phrase (word style)	
W	1-4		4 phonetically rich words	

Table 5: FDB/MDB corpus codes

Corpus identifier	Item identifier	Corpus contents	
A	1-2	2 application words	
B	1	1 sequence of 10 isolated digits	
C	3	1 credit card number (16 digits)	2 connected digits
C	4	1 PIN code (6 digits)	
L	1	1 fixed spelled "forename surname" (set 150)	2 spelled words (letter sequences)
L	2-3	2 spelled "names/words" (set of 40)	
O	1	1 fixed "forename surname"	
O	7-8	2 "forename surname" (set of 10 "full" names)	
S	0-3	4 common sentences	10 phonetically rich sentences
S	4-9	6 speaker specific sentences	

Table 6: SDB corpus codes

Exact specifications of the contents of each item are given in SD1.1.1 (FDB) [1], SD1.1.2 (MDB) [2], and SD1.1.3 (SDB) [3].

It will be checked if all mandatory items are present in sufficient quantities (see section 3.2 below). Further checks will be directed towards:

- are all digits and numbers in the prompt text of I1-2,C1-4, B1, M1, N1 in numerical format
- are all digits present in B1
- is distribution of digits in I1 uniform
- are credit card numbers (C3) and PIN-codes (C4) from set of 150
- are formats of numbers correct (C1-4, B1, M1, N1)
- are all 30 application words present and in sufficient quantities (the minimum number of examples of each word should amount to #speakers/10, for FDB & MDB)
- how many times does each letter occur in spelled items (L1-3)
  (there is no criterion for this; the check is only informative; uniform distribution is recommended)
- are all phones present in the phonetically rich sentences and in sufficient quantities (the minimum number of occurrences for each phone being #speakers/10, for FDB & MDB)
- are all time words present
- are all date words present
- are read city names from set of 500 (set of 25 for 250 speakers MDB)
- are read company/agency names from set of 500 (set of 25 for 250 speakers MDB)
- are SDB fore- & surnames from set of 150, and O7/8 from set of 10 for each speaker

3.2 Validation of missing items

Databases that do not fulfil the following requirement will be rejected:

- At least 95% of the files of each mandatory item (corpus code) must be present (FDB/MDB).

- If one or more calls out of four are missing, then all four calls are considered missing (250 speakers MDB only)
- All data (100%) of at least 120 speakers must be present (SDB).

As missing files are counted: absent files, and files containing only non-speech (i.e., symbols between square brackets) according to the transcriptions. Files with only corrupted speech (i.e., each word in the transcription has *, ~ or &) will be counted and reported, but the statistics on corrupted speech are not used to approve or reject a database. There will be no further automatic comparison of prompt and transcription text in order to decide if a file is effectively missing. In addition, a manual check on 5% of the transcriptions will be carried out (see section 9).

* (mispronunciations), ** (not understandable speech), and ~ (truncations) are counted in the transcriptions of the short items (to be specified in section 9.1) to get an idea of probably useless data. This will not be used to reject or approve a database but it will be supplied as supplementary information in the validation report.

4. ACOUSTIC QUALITY OF THE SPEECH FILES

The following acoustic measurements are performed on each speech file of a database: file length, mean sample value, clipping rate, and SNR value. These measurements are carried out by each individual partner, using SPEX software. The results are passed on to SPEX (as file <database>\DOC\SAMPSTAT.TXT), together with the database to be validated. SPEX will summarise the results of these acoustic measurements in the validation report by means of histograms. These histograms are generated both on file level and on directory (call) level.

The histograms are presented in the validation report just as they are and not further interpreted by SPEX. On the basis of these data the user of the database should be able to decide which acoustic quality is still acceptable for the application at hand.

5. ANNOTATION FILES

5.1 General criteria

- Empty label files should not occur
- No line may exceed 80 characters
- Each line must be delimited by <CR><LF>
- All files must contain the same mnemonics

5.2 Mandatory and optional mnemonics

The following mnemonics are mandatory for FDBs:

LHD: SAM, 5.10
DBN: SPEECHDAT_<language>_<database type>
VOL: FIXED1<language code>_<nr>

SES: <session number>
DIR: <full pathname with backslashes and without final backslash>
SRC: <filename of speech file>
CCD: <corpus code = item code>
CRP: <corpus repetition, empty for FDB>
REP: <location of recording equipment>
RED: <recording date, in format DD/Mmm/YYYY>
RET: <recording time, in format HH:MM:SS>
SAM: 8000 < =sampling freq.>
BEG: <begin sample, usually 0>
END: <end sample>
SNB: 1 < =number of bytes per sample>
SBF: <sample byte order, meaningless with single bytes>
SSB: 8 < =number of significant bits per sample>
QNT: A-LAW < =quantisation>
SCD: <speaker code>
SEX: M / F / UNKNOWN
AGE: <in years/unknown>
ACC: <regional accent, place of growing up>
REG: <region of call>
ENV: <environment of call>
LBD:
LBR: <start>, <end>, [gain], [minimum value], [maximum value],
<orthographic prompt>
LBO: <start sample>, [centre sample], <end sample>, <transliteration>
EXT: [needed if line exceeds 80 chars]
ELF: <end label file>

(Field values in square brackets are optional)

The following mnemonics are optional for FDB databases:

TYP: orthographic
TXF: <name of the prompt sheet text file>
CMT: <comment>
NCH: 1 < =number of channels recorded>
ARC: <region or area code of call>
SHT: <sheet number for prompts>
CMP: <compression software used; field should be empty if this mnemonic is used!>
EXP: <labelling expert>
SYS: <labelling system>
DAT: <date of completion of labelling>
SPA: <SAMPA version>
PHM: <telephone model>
NET: <network>
DSC: <discontinuity marker (empty)>
EDU: <education level>
SOC: <Socio Economic Status>
HLT: <health>
TRD: <tiredness>

STR: <subjective stress level>
RCC: <recording conditions code>
SNL: <subjective noise level>
ASS: <assessment code>

Also the optional mnemonics should preferably be used in the label files. If the information is not available, then the field values should remain empty.

For MDB and SDB somewhat different conventions apply:

- For MDB the mnemonic NET is obligatory, and, in addition, CRP is obligatory for the 250 speaker MDB databases. CRP counts the number of occurrences of a specific item value for a speaker that is spoken more than once (e.g. a specific place name). If a speaker misses the word in a call, CRP should not be increased.
- For SDB the mnemonics HLT, TRD,CRP,RCC,SNL,NET, and PHM are obligatory.
- VOL is MOBIL1<language code>_<nr> for MDB and VERIF1<language code>_<nr> for SDB

Restrictions in the order of mnemonics:

- LHD and TYP are first
- LBR and LBO come after LBD
- ELF is end of file keyword

If EXT is used for lines longer than 80 characters, then take care that line interruptions are only allowed between words, not within words!

For spontaneous speech LBR should contain a mnemonic word. See SD1.3.1 section 5.2 [7])

5.3 Transliterations

The following criteria are valid for the orthographic transcriptions:

- The transliterations are case-sensitive unless specified otherwise in the documentation
- Punctuation marks should not be used in the transliterations
- Digits and numbers must appear in full orthographic form
- In principle only the following symbols are allowed to indicate non-speech acoustic events: [fil] [spk] [sta] [int].
- Asterisks should be used to indicate mispronunciations
- Double asterisks should be used for not understandable parts
- Tildes should be used to indicate recording truncations (and can therefore only appear at the beginning and/or at the end of the utterance)
- Ampersand should be used for typical GSM distortions (MDB and SDB)

6. LEXICON

For the lexicon (in <database>\TABLE\LEXICON.TBL) the following checks are carried out:

- Each line is delimited by <CR><LF>
- The entries should be alphabetically ordered
- Only SAMPA symbols are used in phonemic transcriptions
- Phoneme symbols are separated by blanks
- A line in the lexicon should have the following format:
  <grapheme form> <TAB> [<frequency> <TAB>] <phoneme transcription> [<alternative phoneme transcriptions separated by TABs>]
  (All these fields must appear on a single line)
- Words with *, &, or ~ may not appear in the lexicon
- The orthographic lexicon entries should exactly match the transcriptions

Frequency information is optional. Also alternative transcriptions are optional. They may follow the first transcription, separated by [TAB] or have a separate entry (in case also frequency information is supplied).

Optional information that may be present in the phonemic transcriptions include: stress, word/morphological/syllabic boundaries. If provided, then the symbols used should follow the SpeechDat conventions defined in SD1.3.2 [8].

Orthographic entries are as a rule split by spaces only, not by apostrophes, and not by hyphens.

The lexicon should be complete. A check is carried out on the transcriptions in the LBO fields in the label files in order to find out if the lexicon is undercomplete or overcomplete. Undercompleteness implies rejection of the database, overcompleteness does not.

The lexicon validation is focused on the format of the lexicon table only; the lexicon contents (i.e. the correctness of the phonemic transcriptions) are not validated.

7. SPEAKER INFORMATION AND DISTRIBUTION

7.1 Format specifications

As for the format of the speaker tables the following requirements are set:

- Each line should end with <CR><LF>
- Between field values [TAB]s are used

7.1.1 FDB and MDB databases

For FDBs and MDBs the obligatory fields in SPEAKER.TBL are:

1. Unique number (speaker/caller)
2. Speaker sex
3. Speaker age
4. Accent

The speaker code must be unique. If not, SESSION.TBL must be supplied which contains the speaker information using the session number as key instead of the speaker code (see also section 8.2 of this document, and section 6.3 of SD1.3.1 [7]).

Optional speaker information is:

- height
- weight
- native language
- ethnic group
- education level
- smoking habits
- pathologies
- socio-economic status
- accent
- health
- tiredness

7.1.2 SDB databases

For SDB databases both SPEAKER.TBL and SESSION.TBL are mandatory files. The following fields are mandatory in SPEAKER.TBL:

1. Unique number (speaker/caller) (mnemonic SCD)
2. Speaker sex (SEX)
3. Speaker age (AGE or DOB)
4. Accent; region of growing up (ACC)

And the following fields should be in the SESSION.TBL:

1. Session number (SES)
2. Speaker code (SCD)
3. Health (HLT)
4. Tiredness (TRD)

A recommended category is:

5. stress (STR)

A full overview of the information to be provided in the SESSION.TBL file is given in section 8.2.

For SDBs an additional set of speaker list files is obligatory. These files are in the INDEX subdirectory, and have the extension LST. The speaker list files describe all label files belonging to a specific speaker. Their names obey the following template:

```
DD SSSS LL . LST
```

were `DD` and `LL` are the database and the language codes already defined, and `SSSS` is the four-digit speaker codes used in the SDB recordings (see SD1.3.1, section 7.3 [7]).

7.2 Balances of sex, age and accent

The misbalance of sexes may be 5% at maximum. This means that the proportion of calls from male and female speakers must be in the interval 45-55% for both sexes.

For speaker ages the following criteria are valid:

Age interval:	Proportion:	Requirement:
<16	>= 1%	Recommended
16-30	>= 20%	Mandatory
31-45	>= 20%	Mandatory
46-60	>= 15%	Mandatory

Table 7: Required age distribution

The age criteria are meant for the whole database; they need not to apply, in a more strict sense, for male and female speakers separately.

For SDB the figures in table 7 are strongly recommended but not obligatory.

The balance of regions is validated by checking the speaker file and counting how many speakers called from which region. The result is then compared to the information in the database documentation (DESIGN.DOC). For FDBs all accent regions should be represented by at least 0.5% of the speakers in the database [4] .

For SDBs the following requirements were set up:
- 120 speakers should be included who call 20 times each.
- an interval of three days or more is demanded between consecutive calls
- speakers from at least two dialect regions should be recorded
- the minimum number of speakers per dialect region is eight

8. RECORDING CONDITIONS

8.1 The REC_COND.TBL file

The recording platform should be specified in the documentation of the database (DESIGN.DOC). Information about the recording process is contained in the label files,

e.g. recording date, recording time, recording place, regions of call, calling environment, telephone network and handset used. This information and some other information can also be stored in the recording condition table. This file is obligatory for SDBs and optional for FDBs and MDBs. This file should meet the following requirements:

- It should be stored as <database>\TABLE\REC_COND.TBL
- Each line must be terminated by <CR><LF>
- The first field must be the recordings condition code (mnemonic RCC) which must be also used in the label files
- Fields are separated by TABs

The following information fields must be provided in REC_COND.TBL:

1. unique recording conditions code (mnemonic RCC)
2. region of call (mnemonic REG)
3. environment (mnemonic ENV)

Optional information includes:

1. telephone model (mnemonic PHM)
2. telephone network (mnemonic NET)
3. telephone area code (mnemonic ARC)
4. subjective noise level (mnemonic SNL)

NET is mandatory for MDBs, and NET, PHM and SNL are mandatory for SDBs.

8.2 The SESSION.TBL file

Recording condition information can also be stored in the SESSION.TBL file. This file is mandatory for SDB. For FDB and MDB it can be supplied instead of SPEAKER.TBL, as mentioned in section 7.1.1. It should contain the following items:

mnem.	FDB	MDB	SDB	comment
SES:	Yes	Yes	Yes	session code
RED:	Yes	Yes	Yes	recording date
RET:	Yes	Yes	Yes	recording time
SCD:	-	-	Yes	speaker code[2]
AGE:	Yes	Yes	*these fields*	speaker age
SEX:	Yes	Yes	*must be in*	speaker sex
ACC:	Yes	Yes	*SPEAKER.TBL*	speaker accent
HLT:	-	-	Yes	health
TRD:	-	-	Yes	tiredness
STR:	-	-	Recommended	stress
RCC:	-	-	-	recording condition code[2]
REG:	Yes	Yes	Yes	calling region

[2]If speaker and recording condition tables are supplied, the speaker code and recording condition codes must be unique. In any case these codes can be put in this table even if not unique.

ENV:	Yes	Yes	Yes	calling environment
NET:	-	Yes	Yes	telephone network
PHM:	-	-	Yes	hand-set type used
SNL:	-	-	Yes	subjective noise level

Table 8: Contents of SESSION.TBL per database type (copied from SD1.3.1).

8.3 FDB requirements

For FDBs at least 2% of the calls must be made from a public place. This will be validated by checking the ENV mnemonic in the label files.

8.4 MDB requirements

Whether or not a recordings conditions table is provided, for MDBs it will be checked (from the label files) if recordings are made from the four prescribed environments in equal distribution:

1. Home/office (ENV-value: HOME-OFFICE)
2. Public place (background talking) (ENV-value: PUBLIC_PLACE)
3. Pedestrian by road side (traffic emission noise) (ENV-value: STREET)
4. Passenger in moving car, bus (traffic immision noise) (ENV-value: VEHICLE)

For the 250 speakers' MDBs there is no tolerance for deviations; all 250 speakers have to call from all four environments. For the 1000 speakers' MDBs a deviation of 5% from 25% is tolerated, i.e. each environment should be represented by a minimum of 200 out of 1000 calls (and a maximum of 300 calls).

At least 90% of the calls should be made via the GSM network.

8.5 SDB requirements

For SDBs some other requirements are tested:

- 70% of the calls of each speaker is made in a quiet environment (HOME, OFFICE, CAR STOPPED);
- 30% of the calls of each speaker is made in a noisy environment (PUBLIC PLACE, STREET, MOVING VEHICLE);
- 50% of the calls of each speaker is made over the fixed network ;
- 50% of the calls of each speaker is made over the GSM network.

Validation criteria for the <u>environment</u> are:

1. 10% of the callers (=12) are permitted to call from 1 wrong environment
2. 2.5% of the callers (=3) are permitted to call from 2 wrong environments

Validation criteria for the <u>network</u> are:

1. 10% of the callers (=12) are permitted to make 1 wrong network call

2. 2.5% of the callers (=3) are permitted to make 2 wrong network calls

9. TRANSCRIPTION

9.1 Type of errors

Two types of errors are distinguished:

1. Errors in the transcription of speech
2. Errors in the transcription of non-speech (background noises)

Errors in the transcription of truncations, mispronunciations, word fragments and not-understandable fragments are counted as errors in the transcription of speech. Only errors in the transcription of non-speech acoustic events (i.e., in [fil], [spk], [sta], and [int]) are counted as non-speech errors.

The transcription validation is carried out by a trained native speaker of the language concerned. The transcriptions in the label files are checked by listening to the corresponding speech files and correcting the transcriptions if necessary. As a general rule it is maintained that the delivered transcription should always have the benefit of the doubt and that only overt errors should be corrected. A subdivision is made in long items and short items.

Short items are:
- isolated digit
- time phrases
- date phrases
- yes/no questions
- names
- application words
- phonetically rich words

Long items are:
- isolated digit string
- connected digits
- natural numbers
- money amounts
- spelled words
- application phrases
- phonetically rich sentences

The validation is carried out by taking 5% of the mandatory short items and 5% of the mandatory long items in the first 1000 speakers of an FDB4000+ corpus. This amounts to 1150 short items and 1000 long items. For the other databases the same number of items is selected for transcription validation.

9.2 Criteria for validation

The main criteria for the validation of the transcriptions are:

- For speech a maximum of 5% of the validated items (=files) may contain a transcription error.
- For non-speech a maximum of 20% of the validated items (=files) may contain a transcription error.

All non-speech symbols are mapped onto one during validation, i.e. if a non-speech symbol was at the proper location then it is validated as correct, regardless if it is the *correct* non-speech symbol or not.
Further, only noise *deletions* in the transcription are counted as wrong, not noise insertions.
The error percentage is only determined on item level, not on word level.

9.3 Statistical reliability

As was already pointed out, 1150 short items and 1000 long items are checked for all databases. We computed confidence intervals for the errors in all the transcriptions in the database based on the error percentage found in a sample of this size. Thus, we computed the confidence intervals at 95% reliability for an error percentage of 5%, 50% and 95%, respectively. The results are presented in table 9.

Error percentage		
	long items	**short items**
5%	3.6% - 6.4%	3.7% - 6.3%
50%	46.9% - 53.1%	47.1% - 52.9%
95%	93.6% - 96.4%	93.7% - 96.3%

Table 9: 95% Confidence intervals for a 5% sample containing 1150 short items and 1000 long items.

9.4 Spelling check

A formal spelling check will not be carried out by SPEX. It is recommended that partners report the results of a spelling check that they carried out themselves in the documentation of the database.

10. REFERENCES

[1] R. Winski: *Definition of corpus, scripts and standards for Fixed Networks*. SpeechDat Technical Report SD1.1.1., 1996.

[2] J.G. van Velden, D. Langmann & M. Pawlewski*: Specification of speech data collection over mobile telephone networks*. SpeechDat Technical Report SD1.1.2., 1996.

[3] K. Kordi: *Definition of corpus, scripts and standards for Speaker Verification*. SpeechDat Technical Report SD1.1.3., 1996.

[4] F. Senia et al: *Environmental and speaker specific coverage for fixed networks*. SpeechDat Technical Report SD1.2.1, 1996.

[5] A. Nataf: *Environmental and speaker specific coverage for SDB*. SpeechDat Technical Report SD1.2.3., 1996.

[6] H. van den Heuvel & E.P. Sanders*: The validation of SpeechDat(M) databases: results and recommendations*. SpeechDat Technical Report SD1.3.0., 1996.

[7] F. Senia: *Specification of speech database interchange format*. SpeechDat Technical Report SD1.3.1, 1996.

[8] F. Senia & J.G. van Velden: *Specification of orthographic transcription and lexicon conventions*. SpeechDat Technical Report SD1.3.2, 1996.