DELIVERABLE IDENTIFICATION

Identification number	ED1.4.2
Type	Technical Report
Title	Validation criteria
Status	Final
Deliverable	ED1.4.2
Work Package	WP 1
Task	Task 1.4
Period covered	T01-T04
Date	27/10/1999
Version	1.3
Number of pages	17
Author(s)	Henk van den Heuvel (SPEX)
Workpackage (WP)/ Task (T) responsible	WP 1 – Valery Galounov (AudiTech) Task 1.4 – Henk van den Heuvel
Project contact point:	Jerome Boudy, Matra Nortel Communications Rue J.P. Timbaud – 78392 Bois d'Arcy – France Phone: +33 1 34 60 , Fax: +33 1 34 60 88 32 E-mail: Jerome.Boudy@matra-com.fr
CEC project officer	Mr. M. Ljungquist
Status	Public
Actual distribution	Consortium and CEC
Supplementary notes	

Key words	Speech, database, transcription, orthography, Recognition.
Abstract	This document provides a specification of the validation criteria employed for the telephone speech databases to be collected in the SpeechDat(E) project. These criteria are based on the validation specifications used in SpeechDat(II) and SALA slightly modified according to several workshop discussions. Deviations from the original SpeechDat criteria are explicitly indicated in a concluding section.
Status of the abstract	Public

Received on	
Recipient's catalogue number	

DOCUMENT EVOLUTION

Version	**Date**	**Status**	**Notes**
1.0	11/05/1999	First draft	
1.1	26/05/99	Draft	Updated after discussions in St. Petersburg, May 1999
1.2	04/10/99		Updated after prevalidations and discussion in Wroclaw. Discussion on speaker accent is included in Section 7.
1.3	27/10/99	Final	Final outcome on speaker accent included in section 7.

Contents

INTRODUCTION

Comment:

The objective of this document is to make explicit the criteria that SpeechDat-East 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 evolved from experiences with the SpeechDat(II) project and additional discussions in the SpeechDat(E) project group. It was decided to stick as close as possible to the criteria laid down in SD1.3.3 of the SpeechDat(II) project [5].

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

1 DOCUMENTATION
2 DATABASE STRUCTURE, CONTENTS AND FILE NAMES
3 DATABASE ITEMS AND COMPLETENESS
4 SAMPLED DATA FILES
5 ANNOTATION FILES
6 LEXICON
7 SPEAKERS
8 ENVIRONMENT CONDITIONS
9 TRANSCRIPTION
10 TRAINING / TEST SPECIFICATIONS

As section 11 we add a summary of validation criteria which deviate from those in SpeechDat(II).

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;
- 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;
- analysis of frequency of occurrence of the phones represented in the phonetically rich sentences and in the phonetically rich words, and in the full database; the last count should be made on all read items in the database
- for rare phones which have a lower frequency than allowed as minimum, a listing and motivation should be given; SPEX may consult external experts for verification of these motiviations; also possible mappings with other phones should be given
- the prompting design (e.g. how items were spread to prevent list effects);
- recording, and telephone link (net) description;
- speaker demographics information:
  - sexes: males, females, how many of each;
  - regions, which and how many speakers of each;
  - age groups, how many speakers of each;
- recording conditions;

- annotation information:
  - procedure used;
  - quality assurance;
  - standard character set used for transcription (ISO-8859-?);
  - spelling standard used;
  - 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;
  - list of symbols used to denote mispronunciations and interrupts;
- lexicon information:
  - procedure followed to obtain phonemic forms from orthographic input,
  - format of the lexicon
  - case-sensitivity of orthographic entries,
  - selection, sorting and case of the entries,
  - phone set used (SAMPA),
  - information captured in the phone transcriptions (assimilation and reduction rules),
- any other language-dependent information or conventions;
- reference to the validation report VALREP.TXT;
- 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 is distributed among partners by SPEX.

Language-dependent information may be delivered in a file in PDF or PS format, in order to cope with language-specific characters, not generally supported by WORD.

2 DATABASE STRUCTURE, FILE NAMES, AND CONTENTS

Checks will be directed towards:

- All obligatory files are present (see SD1.3.1, section 9 and below)
- Directory structure and filenames are correct. The ISO-639 standard is used for the language code.
- All text files are in MS-DOS format (<CR><LF> at line ends)
- The following files should be in \<database_name>\DOC:
  . DESIGN.DOC
  . TRANSCRIP.DOC (recommended)
  . SAMPALEX.PS
  . ISO8859<nr>.PS
  . SUMMARY.TXT
  . SAMPSTAT.TXT
- Tables should be in \<database_name>\TABLE
. LEXICON.TBL
. SPEAKER.TBL or SESSION.TBL
- A README.TXT file should be in the root describing all files on the CD-ROM.
- A file containing a shortened version of the volume name (11 chars max.)

should be in the root directory. The name of this file is DISK.ID.
This file supplies the volume label to UNIX systems that cannot read
the physical volume label.
- Each table file and index file has a header line explaining the columns
- A copyright statement is present
- The contents list (CONTENTS.LST) is in the INDEX directory and has correct format & contents
- The summary file (SUMMARY.TXT) is in the DOC directory and has correct format & contents
- Contents lowest level subdirectories should be of one call only
- Empty (i.e. zero-length) files are not permitted
- File match:
For each label file there must be one speech file and vice versa.
-Test set file must be included in INDEX directory (see section 10)

See below a table of all obligatory files.

File(s)	Obligatory
Speech files	Yes
Label files	Yes
Speaker information file	Yes, one or both of these
Session information file	
Recording condition file	-
Lexicon file	Yes
Contents index file	Yes
Test set file	Yes
Corpus contents index files	-
Speaker list files	-
Readme file	Yes
Disk id file	Yes
Copyright file	Yes
Summary file	Yes
Design documentation	Yes
Transcription manual	Recommended
ISO-8859 table	Yes
SAMPA table	Yes
Alternate spelling list	-
Source files	-
Prompt files	-

3 DATABASE ITEMS AND COMPLETENESS

3.1 Mandatory items

The mandatory database items are listed in ED1.12.1-5.

Checks will be directed towards:

- Checks at ***prompt*** level :
  - credit card number: set of 150;
  - PIN code if used : set of 150;
  - Application words: set of 25-30 (33 for Russian);
  - city names: set of at least 500;
  - company/agency names: set of at least 500;
  - forename/surname set of 150;
  - surname set of 150.
  - phonetically rich sentences: Minimum number of different sentences is 2250 for Russian and 1200 for the others
  - phonetically rich words: Minimum number of different words is 2000 for Russian and 800 for the others

Formats and ranges of connected digit strings and numbers:
- Natural numbers may be larger than 1,000,000, but should contain a maximum of 4 significant digits (counted from the left side)
- Digits only in numerical format in prompts of digits and numbers;
- Credit card number should contain 14-16 digits in blocks of max. 4;
- PIN code should contain 6 digits;
- B1 should contain every digit

- Checks **from tables in DESIGN.DOC** :

- Phon. rich sentences: Min. frequency of each phoneme: #calls/10
- Phon. rich words : Min. frequency of each phoneme: #calls/10

4.2 Validation of missing items

It will be checked if all mandatory items are present in sufficient quantities.
Databases that do not fulfil the following requirement will be rejected:

- A maximum of 5% of the files of each mandatory item (corpus code) may be effectively missing;
- Another maximum of 5% of the phonetically rich sentences may contain corrupted speech only;

As *effectively missing* files are counted: absent files, and files containing only non-speech (i.e., noise symbols between square brackets) according to the transcriptions. Files with only *corrupted speech* are files for which *each* word is distorted in some way according to the transcription. In addition, a manual check on the transcriptions of 2000 utterances will be carried out (see section 9).

* (mispronunciations), ** (not understandable speech), % (GSM distortions), 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.

Items within the following homogeneous categories of corpus codes may compensate for each other in order to meet the completeness criteria:

- A1-6
- C1-4
- B1, I1
- D1-2
- L1-3
- O2-3
- O8-7
- S0-9, Z0-1
- T1-2
- W1-4

4 ACOUSTIC QUALITY OF THE SPEECH FILES

The speech files of the databases must be delivered as A-law.

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 any specific application at hand.

Calls with a very low average SNR and very high average clipping rate will be reported and inspected by SPEX, which may result in a recommendation not to use the call for training (and testing).

5 ANNOTATION FILES

5.1 General criteria

- Empty label files should not occur
- Each line must be delimited by <CR><LF>
- All files must contain the same mnemonics
- All obligatory mnemonics (see ED1.3 [7]) must be used
- The mnemonic NET is obligatory if mobile calls are included in the recordings
- No illegal field values are allowed
- Each recording session directory should not refer to multiple sheet IDs

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
- 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)

- The first line is a header line containing a description of the record fields
- Words with *, %, or ~ may not appear in the lexicon
- The orthographic lexicon entries should exactly match the transcriptions
- The correctness of the phone transcriptions as such is not checked
- All and only SAMPA symbols are used in phonemic transcriptions (or agreed extensions)
- If optional information is encoded (stress, boundaries of morphemes or syllables) the SpeechDat conventions should be followed
- All words in the LBO transcriptions are in the lexicon (the lexicon should be complete)
- A rough check on the plausibility of phone transcriptions is carried out, but not by a native speaker of the language

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).

The correct format of the lexicon is described in [7].

The lexicon should be complete. The completeness 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.

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

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

Optional speaker information is:

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

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

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.

The balance of regional accents is validated by checking the label files and counting how many speakers belong to which dialect region. In most countries, ACC (being the place where people grew up during their primary school period) will bear the closest relation to the pronunciation variant which a speaker will entail, but this need not be. Both ACC or REG may be used as the criterion to recruit speakers, as long as the

regions themselves are defined on the basis of dialectal motivations. For validation ACC or REG will be used accordingly, based on the regional recruitment strategy outlined in the DESIGN.DOC file. The distribution of speakers over the dialect regions should be proportional to that of the population with a deviation of 5% at the maximum and a minimum representation of 5% of the calls for each dialect region. The distribution resulting from the label files is compared to the information given in the DESIGN.DOC file.

8 RECORDING CONDITIONS

- The REC_COND.TBL is optional. But if supplied it should have the correct format and contents (see [7]).

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.

- 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;
- If calls from mobile networks are included, then the attribute NET is compulsory in the label files, and the maximum of such mobile network calls is 5%.

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 exact transcription conventions are in ED1.4.1.

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 1000 short items and 1000 long items.

9.2 Criteria for validation

Automatically tested for each label file are:

- 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)
- Percent sign should be used for typical GSM distortions

The criteria for the manual validation of the transcriptions by the native speaker 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 *(except [sta])* 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, 1000 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 below.

Error percentage	Confidence interval
5%	3.6% - 6.4%
50%	46.9% - 53.1%
95%	93.6% - 96.4%

For the whole sample of 2000 utterances the 95% confidence intervals are:

Error percentage	Confidence interval
5%	4.0% - 6.0%
50%	47.8% - 52.2%
95%	94.0% - 96.0%

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 TRAINING /TEST SET SPECIFICATIONS

The following criteria are taken from SD1.3.4 [6]:

- Each database should contain an index file with selected test sessions (500 sessions for the Russian database and 200 sessions for the other databases)
- The existence of these selected test sessions in the database is checked
- If tests are specified, then it is checked whether the correct formats are used (mnemonics, field values, files' existence)

11 DEVIATIONS FROM SPEECHDAT(II)

The following lists all validation criteria that deviate from the SpeechDat(II) criteria.

- The DESIGN.DOC file should also contain a table of phone counts for the full database;
- An additional maximum of 5% of the phonetically rich sentences may contain corrupted speech only (additional criterion, section 3.2);
- City names and agency names should each come from a list of *at least* 500 names;

- Natural numbers may exceed 1 Million, provided they do not contain more than 4 significant digits;
- Phonetically rich sentences: each unique sentence should not appear more than 10 times;
- Phon. rich words : Min. frequency of each phoneme: #calls/10 (relaxation, section 3);
- Phon. rich words : each unique word should not appear more than 5 times;
- Calls with a very low average SNR or a very high average clipping rate will be reported and inspected by SPEX, which may result in a recommendation not to use the call for training (and testing). (additional, section 4);
- Line lengths in label files may exceed 80 characters;
- The distribution of the dialect regions among the calls should be proportional to that of the population with a deviation of 5% at the maximum, and a minimum representation of 5% of the calls for each dialect region;
- REG may represent speaker accent in specific cases;
- If calls from mobile networks are included, then the attribute NET is compulsory in the label files (section 5), and the maximum of such mobile network calls is 5% (section 8);
- During transcription validation, the [sta] symbol will not be treated as equivalent to other symbols between square brackets (additional criterion, section 9.2);
- The sample of short items for transcription validation is reduced from 1150 to 1000 (section 9.1).

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

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

[5] H. van den Heuvel: *Validation criteria*. SpeechDat Technical Report SD1.3.3, version 1.9, 1997.

[6] G. Chollet, B. Lindberg, F.T. Johansen, F. Senia: *Test definition and specification*. SpeechDat Technical Report SD1.3.4, 1997.

[7] Jan Černocký, Petr Pollák: *Specification of speech database interchange format*. SpeechDat-East Technical Report ED1.3., 1999.

13 APPENDIX: SPEECHDAT REVALIDATION PROCEDURE

COMMENT: Document made for SpeechDat(II) and updated for SpeechDat(E).

1. PROBLEM OUTLINE

The validation centre, SPEX, is in charge of the validation of the databases generated in the SpeechDat project. A database is delivered and its format and contents are checked by SPEX against the specifications defined in a set of deliverables written in WP1. After this validation a report VALREP.TXT is written which contains the results of this validation.

This document may report on shortcomings of the database which are difficult to repair (e.g. the absence of obligatory items which were not included in the prompt sheets) and on shortcomings which can easily be mended (e.g. typo in label file mnemonic).

Partners will in most cases consider it desirable to correct some of the reported flaws in their database, especially the easy to fix ones, and put the rectified files on the final CDs. However, in the VALREP.TXT, which should also be put onto the CDs, these (meanwhile obsolete) comments would still be present.
SPEX considers it an undesirable situation a. if a database is printed with an obsolete validation report, and b. if a database is not rechecked after corrections were made by the producer.

In other words, a careful but feasible revalidation protocol should be described.

2. POINTS OF DEPARTURE

SPEX validates a database and the report is discussed in the Consortium. In case of a problematic database the Consortium makes a proposal for database modification to the database producers/owners.
The final decision as to acceptance/rejection of a database is with the Consortium.

The owner of a database is responsible for the database and its contents;
SPEX is responsible for the validation report.
SPEX will not carry out any modifications of the database;
The database owners will not modify the validation report.

Therefore, the corrected database, or parts of it, should be re-sent to SPEX for a recheck. This recheck will be termed revalidation.

The database offered for revalidation will be put at the end of the queue. It must wait until all other databases that wait for their first validation are ready, and until all other databases that were offered for revalidation earlier are ready. - Only when a free time slot appears (due to e.g. delayed deliveries of other databases) can revalidation databases be dealt with earlier.
The database owner pays for the additional validation (see section 5).

3. WHICH DATABASES WILL BE REVALIDATED?

After the first validation there will be four categories of databases:

1. Databases that are accepted and fulfil all
   requirements (of course, this will be the largest group!)
2. Databases that are accepted anyhow but that
   contain some minor shortcomings which can be repaired
3. Databases that are only accepted if some
   shortcomings are repaired
4. Databases which are rejected right out

Databases in categories 2 and 3 will be accepted for revalidation.
This means that a database which is already approved (category 2)
can be offered for revalidation. This is to avoid that corrected databases
from category 3 may end up in a better condition than those from category 2.
Databases from category 2 must be revalidated, when rectifications are made,
no matter how minor they are. In this case the database owner should decide
whether he thinks it is worthwile to revalidate and accept the queueing
time and additional costs, or to leave things as they are.

In case of very minor changes, SPEX may accept that a partner
includes an extra section in DESIGN.DOC listing these modifications and
stating explicitly that these modifications were carried out after the
validation was carried out and are for that reason not included in the
validation report. In any case, SPEX has to agree with such a procedure.

Databases in category 3 must be revalidated anyhow.

4. WHAT CAN BE REVALIDATED?

In principle each part of the database may be revalidated, also the
transcription validation, as long as the partner accepts the additional
costs and the queueing time involved.

The validation report makes clear which part of the database
is subject to revalidation. On this basis it can be discussed with the
database producer if the full database should be delivered (on CDs)
or only parts (via the FTP site).

5. REVALIDATION COSTS

Revalidation costs will be paid by the database owner.

For category 2 databases (see section 3 above) SPEX will make a
price offer when the partner concerned gives an overview of the items
for which revalidation is requested.
If the revalidation work is minor and can be completed in
-say- less than half a working day, SPEX will not charge the database owner,
since then the overhead incurred by the billing
process might be higher than the total amount of the bill.

For category 3 databases SPEX will make an overview of the items that must be
revalidated and of the costs involved.

Revalidation costs can be paid from undeclared project money if the database
producer has any left. Otherwise, the payment has to come from the partner's
own resources.

The maximum costs for a full revalidation are 4 kEUR. The composition of these costs is about as follows:

- data transport from CD to disk: 0.2 kEUR
- database design and (effectively) missing files: 0.8 kEUR
- transcription validation: 0.8 kEUR
- lexicon (SAMPA symbols & completeness): 0.4 kEUR
- use of mnemonics and values in label files: 0.4 kEUR
- documentation: 0.2 kEUR
- database structure and filenames: 0.2 kEUR
- acoustical measurements (given SAMPSTAT.TXT): 0.1 kEUR
- environment check: 0.1 kEUR
- administration and overhead: 0.6 kEUR