DELIVERABLE IDENTIFICATION

Identification number	LE4-8334 - D1.3.1
Type	Technical Report
Title	Validation Criteria
Status	Final
Deliverable	D1.3.1
Work Package	WP 1
Task	Task 1.3
Period covered	T01-03
Date	12/09/2000
Version	3.5
Number of pages	21
Author	Henk van den Heuvel, SPEX
Workpackage (WP)/ Task (T) responsible	WP 1 – Luis Arevalo, Stephan Euler (BOSCH) Task 1.3 - Henk van den Heuvel (SPEX)
Project contact point:	Robrecht Comeyne L&H Speech Products Flanders Language Valley 50 B-8900 Ieper Belgium Tel. +32 57 22 88 88
CEC project officer	Mr. M. Ljungquist
Status	Public
Actual distribution	Consortium and CEC
Supplementary notes	

Key words	Telephone, speech, database, validation
Abstract	This report presents an account of the criteria that databases produced in the framework of SpeechDat-Car should meet in order to be accepted as valid and equivalent databases. These criteria involve: 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 QUALITY Also the validation procedures and protocols are described in this report.
Status of the abstract	Public

Received on	
Recipient's catalogue number	

DOCUMENT EVOLUTION

Version	Date	Status	Notes
1.0	29/04/98	First draft	Only validation criteria addressed
2.0	18/06/98	Draft	Doc. Confined to Validation criteria, and updated after Munich workshop
3.0	27/06/98	Draft	Updated after Paris workshop
3.1	31/07/98	Prefinal	Prefinal version for consortium.
3.2	05/08/98	Prefinal	Prefinal version for Steering Committee
3.3	30/03/99	Final	Finalisation after 6th workshop in Nijmegen
3.4	22/10/99	Final	Finalisation after database prevalidations. Final criteria for acoustic quality (section 5) inserted.
3.5	13/10/2000		Modifications after validation of first full db.

CONTENTS

INTRODUCTION

Comment:

The objective of this document is to make explicit the criteria that SpeechDat-Car 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 other (mainly SpeechDat(II)) databases and discussions amongst SpeechDat-Car partners. The validation criteria developed within the SpeechDat(II) project were used as a starting point [1]. The validation criteria outlined in the present document were discussed in Munich (May 1998), and Paris (June 1998). The decisions made there are included in the present document. Also further updates discussed in Nijmegen (March 1999) are included and modifications that turned out to be necessary after the validation of the first full databases..

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 or a reference to these specifications is contained in the present document, as their fulfilment is of immediate importance for the acceptability of a database. Details will have to be looked up in the deliverables concerned, the most relevant of which are D1.12 [2] on database contents, D1.3.3 [4]. on format specifications, and D1.3.2 [3] on the transcription conventions. It is important to note that the specifications in the aforesaid 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 QUALITY

The criteria outlined in this report will be worked into a precise check list which will serve as the basis for the database validation reports and distributed among the consortium partners.

2. 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, if needed if a phoneme is not sufficiently well represented otherwise, see section 4.3); the phone counts should be made at transcription level, not at prompt level;
- the prompting design (e.g. how items were spread to prevent list effects);
- recording platforms, platform interaction, number and type of microphones and their positioning, synchronisation of simultaneous recordings, and telephone link 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: environments, GSM network, handsets; and the distributions of the environments in the database; statement if the GSM network operator supports/warrants EFR coding;
- annotation information:
  - procedure used;
  - quality assurance;
  - standard character set used for transcription (ISO-8859-1);
  - 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;
- information on test (set) specification(s)
- reference to the validation report VALREP.DOC;
- 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.

3. DATABASE STRUCTURE, FILE NAMES, AND CONTENTS

3.1 Directory names and file names

The database should have the directory structure and file names, as specified in [4]. A summary table of all mandatory files is in section 8 of [4].

3.2 The DOC directory

The following files should be in \<database_name>\DOC:
. DESIGN.DOC
. D14V??.DOC
. TRANSCRP.DOC
. SPELLALT.DOC (optional)
. SAMPALEX.PS
. ISO8859<n>.PS
. SUMMAR{0|1|2|3|Y}.TXT
. SAMPSTA{0|1|2|3|T}.TXT

The validation of the DESIGN.DOC main documentation file is described in section 2. D14V??.DOC contains the most recent platforms specifications. TRANSCRP.DOC should contain the transcription instructions to the transcribers (in the native language and/or in English). ISO8859<n>.PS is a postscript file containing the ISO-8859-<n> character table used for orthographic transcription. The SAMPALEX file lists the SAMPA symbols used for the phonemic transcriptions in the lexicon together with an example. SUMMAR{0|1|2|3|Y}.TXT contains an overview of all items recorded for each session per channel. SAMPSTA{0|1|2|3|T}.TXT is the output per channel of the acoustical check on the speech files performed by each partner. Software for this check will be provided by SPEX.

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

3.3 The TABLE directory

Tables should be in \<database_name>\TABLE
. LEXICON.TBL
. SPEAKER.TBL and SESSION.TBL
. REC_COND.TBL

The validation of LEXICON.TBL is dealt with in section 7; the validation criteria for the SPEAKER.TBL is given in section 8, and for SESSION.TBL and REC_COND.TBL files in section 9.

3.4 Other directories

The root directory should contain the files:
. README.TXT (containing a description of the files in the database)
. README.HTM (optional, with browser access to all documentation directories)

. COPYRIGH.TXT (copyright statement)
. DISK.ID (character string with volume name for UNIX platforms)
These files will be checked for the one CD (containing the non-speech files of the database, see section 11.2) that SPEX receives for (full) validation.

Index files should be in \<database_name>\INDEX.:
. CONTENT0.LST
. CONTENT{1|2|3|S}.LST (optional)
. V1TRN<language code>.SES
. V1TST<language code>.SES

CONTENT0.LST should have the correct format, specified in [4], and contain the correct information for the close talk microphone channels. Optional are the contents lists for the other channels CONTENT{1|2|3|S}.LST. The V1TRN<language code>.SES and the V1TST <language code>.SES contain the train and test set session numbers, respectively.

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

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

An overview of all obligatory files is given in section 8 of [4].

3.5 Other requirements

All text files should be in MS-DOS format. This concerns the label files, the table (.TBL) files, the index (.LST) files, the .TXT files, and the copyright file.

Empty files are illegal. This is of special relevance for speech and label files.

For each label file there should be one corresponding speech file (which contains a multiplex signal for the in-car recordings; and just one signal for the GSM channel), and for each annotated input channel there should be one corresponding label file.

4. DATABASE ITEMS AND COMPLETENESS

4.1 Mandatory items specifications

It will be checked if all mandatory items are recorded. A structurally missing item will lead to rejection of the database.
The mandatory database items are listed in the tables in [2], section 4. Each individual item should meet the specifications in [2], sections 5 and 6.

Checks will be directed towards (***prompt*** level):

- Sizes of sets and links:
  - credit card number: fixed set of 150; all 16 digits;
  - PIN code if used : set of 150;

- Application words: set of 200;
    - Car application words: set of 96;
    - Mobile application words: set of 39;
    - IVR application words: set of 65;
- language-dependent application words: set of 10;
- voice activation key phrases: set of 5;
- city names: set of 150; 53 names from common set;
- forename/surname set of 150;
- L1 will be linked to O1 (spont. forename or surname);
- L2 will be linked to O3 (= one of the city names);
- Min. number of phones per voice activation command: 10;
- Phon. rich sentences: Max. 5 repetitions per sentence; min. 1100 different sentences;
- Phon. rich words: at least 1500 different words

- Formats and ranges of connected digit strings and numbers:
    - Money amounts in EURO's and Cents
    - Natural numbers between 100,000 and 1,000,000
    - Digits only in numerical format in prompts of digits and numbers ;
    - Credit card number should contain 16 digits in blocks of 4;
    - PIN code should contain 6 digits;
    - Sheet code should contain 4 digits or more

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 requirements will be rejected:

- A maximum of 5% of the files of each mandatory item (corpus code) may be *effectively missing*; this maximum is 10% for the GSM channel;
- Another maximum of 5% of the files of each mandatory item (corpus code) may contain *corrupted* speech only;
- A maximum of 10% of the files of isolated word items may show a *mismatch* between prompt and transcription text;
- These criteria are applied to each annotated microphone/input channel

As *effectively missing* files are counted: absent files, and files containing only non-speech (i.e., noise symbols between square brackets and/or **) according to the transcriptions. Files with only *corrupted speech* are files for which *each* word is mispronounced or truncated according to the transcription. (GSM distortions are not regarded as corruptions).

For the isolated word items (especially the application words, which is the main body of the corpus) a further comparison of prompt and transcription is made. In case the word in the prompt does not appear in the transcription (but no speech at all or another or other word(s) instead), then this should be considered as a missing item. A maximum of 10% of the files may be *mismatching* in this way. It is obvious that physically absent files contribute to this count as well.

If the word is present but has a *, ~ or % marker attached to it, then it is *not* considered as a mismatch. The following items are included in this check: 00-66, O3-7, P1-2, W1-4.

* (mispronunciations), ** (not understandable speech), % (GSM distortions), and ~ (truncations) are counted in the transcriptions of the short items (to be specified in section 10.4) 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.

The checks on item completeness rely on a correct transcription of the speech. To verify the transcriptions themselves, a manual check on 2000 of the transcriptions will be carried out in addition (see section 10).

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

- A1,2
- B1, I1-4
- C1-7
- D1-2
- E1-2
- L1-7
- O2-4
- O5-6
- S1-9
- T1-2
- W1-4
- 00-66
- P1,2
- Z0-9

4.3 Validation of missing words/digits

The check on the completeness of each corpus code described above is accompanied by a more close completeness check of individual words items within a corpus.

These checks are carried out *at transcription level*. A word is counted as present if it is in the transcription, even if it is truncated, mispronounced or GSM-distorted (it may still be useful for test purposes in this case). Only if the word is not present in the transcription is it considered as missing.

The relevant checks are shown in the table below.

Item type (corpus)	Min. Samples required	Max samples achievable
Isolated digit	200 per digit	240
10-digits string	500 per digit	600
Application words	175 (for in-car lines) 150 (for GSM line)	201

Lang. dependent application words	100	120
Voice activation keywords	200	240
Spelt letters	500	
Phon. Rich sentences	250*	
Phon. Rich words	100*	
Dates	40 per month name 75 per day name	50 per month name 85 per week day
Relative dates	40 per expression	60 per expression

*compensation by other items in the database is acknowledged, but must be clearly documented in a separate table in DESIGN.DOC. One table should show the number of repetitions achieved if the phon. rich words or sentences alone are counted and another (column in the) table should show the number of repetitions obtained if other items are added. Which these other items are should be made explicit.

5. 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 the validation centre (as files <database>\DOC\SAMPSTA{0|1|2|3|T}.TXT), together with the database to be validated. The validation centre 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. Histograms will be generated for each recording microphone/input channel. 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.

If the average SNR is
- for a full call measured in the close talk microphone <15 dB
- for a full call measured in a far talk microphone < 5 dB,

auditory and visual inspection by the validation centre will follow which may lead to a rejection of the call for training purposes. The same holds for calls with extremely high average clipping rates. Submission of the respective sessions will then be required from the database producer. For these sessions:

- a maximum of 2 of the recordings may contain clearly artificial external noise throughout the recordings (e.g. signal saturation, and other sound/noise components generated by the recording platform itself, and not by the car and the driving environment)
- Furthermore: Each speech file will come with 1 second silence between the initial DTMF tone (co-occurring with the start of the file) and the end of the prompt beep. Deviations up to 100 ms are accepted. A database will be rejected if larger deviations are observed in 5% or more of the inspected speech files.

The aforementioned histograms over the sessions of the *whole* database will also be used to evaluate if the database fulfils the following requirements:

- at least half of the recording sessions must have an average SNR of 15 dB or more for the *close talk* channel;

- at least half of the recording sessions must have an average SNR of 5 dB or more for the each of the *other* channels;

6. ANNOTATION FILES

Checks will be performed as to:

- Correct use of mnemonics and accompanying values (depending on the recording platform)
- Empty label files should not occur
- Each line must be delimited by <CR><LF> (DOS format)

The correct mnemonics and field values are described in [4].

7. LEXICON

For the lexicon table the following checks are carried out:

- Format check
- All and only SAMPA phoneme symbols are used
- The lexicon contains all words in the transcriptions except corrupted words

The format of the lexicon is described in [4].

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.

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.

8. SPEAKER INFORMATION AND DISTRIBUTION

The speaker table file should have the format specified in [4].

- A minimum number of speakers of 300 should be recorded. The maximum number of sessions per speaker is two.

Below we summarise the speaker distribution criteria as given in [2], section 8.1.

Speaker sex:

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 300 speakers the permitted interval is thus 135-165 speakers per sex.

Speaker age:

For speaker ages the following criteria are valid:

Age interval:	**Proportion:**	**Requirement:**
< 15	0%	Mandatory
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.

Speaker accent distribution:

- A database contains a maximum of 6 accent regions.
- Each accent region is represented by at least 50 speakers

It is recommended to use the same accent regions as defined in the SpeechDat(II) project [5] or merge between these.

Speaker balances are validated by checking the label files and counting how many speakers called from each category. The result is then compared to the information in the database documentation (DESIGN.DOC) and the speaker and session tables.

9. RECORDING CONDITIONS

The session table and the recording condition table files should have the format specified in [4].
If labels such as WTC and CEQ (wipers!) are adjusted during a recording session, then this should be reported in the DESIGN.DOC file.

Environments:

Two environments should be selected from a set of 7 possible environments for each speaker as specified in [2], section 8.2, and in [4], section 4.8.5.

Each environment should be represented by at least 10% of the calls.

10. TRANSCRIPTION

10.1 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], [dit] 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 by 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.

10.2 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] [dit] [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)
- The percent symbol should be used for typical GSM distortions

The full description of the relevant transcription conventions can be found in [3a].

These criteria are checked both automatically on the *full* database, and by the native speaker on the *subset* for transcription validation.

10.3 Criteria for validation

The main criteria for the validation of the transcriptions are:

- For speech a maximum of 5% of the validated utterances (=files) may contain a transcription error.
- For non-speech a maximum of 20% of the validated utterances (=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. Only stationary noise may not be confused with another type of noise.
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.

10.4 Statistical reliability

A random sample of 1000 utterances from the long items and 1000 utterances of the short items is checked for each complete database. We computed confidence intervals for the errors in all the transcriptions in the database based on the error percentage found in a sample of size 1000. 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%

Since the 2000 utterances are sampled from the full database, to ascertain a maximum of independence between the items/utterances in the sample.

Short items are:
- isolated digits
- 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
- spontaneous sentences

The 2000 utterances are randomly chosen over the annotated input channels. The number of annotated input channels is typically 1, but this number may be 2 and maximally 4 for some of the databases. In the latter case 2000/4 = 500 utterances per input channel will be checked. The associated 95% confidence intervals are as shown below.

Error percentage	Confidence interval
5%	3.1%-7.0%
50%	45.6%-54.4%
95%	93.1%-97%

10.5 Spelling check

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

11. Validation procedures

A database is validated in at least two stages: prevalidation and validation.

11.1 Prevalidation

Each partner sends a complete minidatabase of 6 sessions (from at least 3 different speakers) to the validation centre. This minidatabase contains all speech and label files and all other files that are required for a normal validation, but, of course, tailored to the speakers included only. The speakers are typically those that were recorded for the task of platform validation. The goals of the pre-validation are:

1. To detect errors in the database design before the main series of recordings start;
2. To stimulate partners to write their database formatting software in an early stage of the project;
3. To stimulate the validation centre to write the validation software in an early stage of the project;
4. To acquire the data for a comparison of simultaneous in-car recordings and GSM recordings in order to establish a set of standard symbols to annotate GSM recordings.

Most of the checks described in the sections 2-10 are performed on this 6-session data. Excluded from these checks are those for speaker and environment balances, the distribution checks on individual words and digits, and the evaluation of the transcription quality by a native speaker (however, the automatic check on transcription symbols is included),

11.2 Validation

For validation the procedure is as follows:

1. The producing partner sends a CD-ROM with all files, except the speech files, to the validation centre (also the label files must be included!);
2. Immediately after reception of the database and therefore prior to validation, SPEX creates a list of 2000 files that will be used for transcription validation. This list will be sent to the partner in the form of a Perl script which can be used to extract the requested files from the full database. These files are copied onto CD and also sent to SPEX. Alternatively, the producing partner may choose to send the CONTENT?.LST files of the annotated channels, to SPEX prior to step 1. In that case step 2 will be done before, or parallel to, step 1.
3. All checks described in the previous sections 2-10 are carried out;

4. The partner may be requested to send some additional sessions for acoustic quality evaluation (section 5);
5. The result of the validation, the validation report VALREP.DOC, is sent to the producing partner.
6. SPEX will ask the producer for clarifications for deviations observed during validation and communicate both deviations and clarifications to the consortium. The consortium decides about the approval of a database. In case SPEX does not find any serious deviations, the database is accepted without voting.

11.3 Re-validation

If the database is not approved by the Steering Committee, or the producing partner wants to add some modifications to the database after it is accepted, then a revalidation by the validation centre should take place. Such a revalidation takes place at extra costs for the producing party, as described in Appendix A.

In case of minor modifications, the validation centre can agree with an extra section in DESIGN.DOC listing the modifications made after the validation report was written. But this should always first be discussed with the validation centre.

11.4 Validation of a first subset of speakers

If the (optional) validation of a first subset (say 150 speakers) is wished then the procedure described in 12.2 is followed. This validation is not a standard validation and should additionally be paid by the producing partner. Exact details of such a validation are given in Appendix A.

12. REFERENCES

[1] Henk van den Heuvel*: Validation criteria*. SpeechDat(II) Technical Report SD1.3.3. Version 1.9, 1997.

[2] Sandra Dufour*: Specification of the car speech database (definition of corpus, scripts and standard), Car environments and speaker coverage*. SpeechDat-Car Technical Report D1.12. 1998.

[3a] Henk van den Heuvel*: Orthographic transcription conventions*. SpeechDat-Car Technical Report D1.3.2a., 1998.

[3b] Henk van den Heuvel*: Orthographic transcription conventions for GSM acoustic events*. SpeechDat-Car Technical Report D1.3.2b., 1998.

[4] Christoph Draxler*: Specification of database interchange format*. SpeechDat-Car Technical Report D1.3.3., 1998.

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

13. APPENDIX A: Agreement on validation of first 150 speakers and revalidation

SUBJECT: Validation, two-step validation and revalidation in SpeechDat-Car
AUTHOR : Henk van den Heuvel
DATE : 20 MAY 1998
VERSION: 2.0 (after discussion in Munich)

COMMENT: To be added to the minutes of the Munich workshop

This is a reply to AP-02, AP-11 and to a less extent AP-18 of the minutes of the workshop on 2 April in Brussels.

AP-02: two step validation for premature data availability to other projects.
AP-11: implication for validation for other project candidates

1. TWO-STEP VALIDATION

It was decided during the workshop in Munich that the extra validation of the first 150 speakers is not mandatory but optional for a database.
This extra validation of 150 speakers can only be carried out at additional costs. The costs depend on what should be validated (see section 3 below).

If part of a database needs early delivery to a third party, and validation of this part is felt necessary,
then the PRL may contact SPEX for validation of the database.
The PRL will be the intermediary and responsible partner for the contacts and contracts with SPEX regarding the subset validation.
Accordingly, the PRL will make a contract with SPEX for the validation and will pay the agreed costs. Payment by the third party to the PRL (if any) is the responsibilty of the PRL.

1. TIME SCHEDULE

Taking into account :

1. that the prevalidation of databases is planned to be finalised at T12;
2. that the validation of the first set of speakers must follow partly in parallel and partly after this milestone;
3. that producing partners have the obligation to finalise recordings of the first 150 speakers at T15 (B1.1 of TA);
4. that partners also involved in VODIS may be ahead of such a scheme and have prevalidation and the recordings of the first 150 speakers earlier (say T10);

it is reasonable that the validation of the first 150 speakers cannot be expected to be completed before T12 for partners that are also involved in VODIS, and not before T18 for other partners.

The principle of processing for validation is First-In-First-Out.

2. VALIDATION CONTENTS

A full validation encompasses checks of:

- documentation files
- database format/structure and file names
- completeness of recordings (structurally and incidentally missing files; completeness of channels)
- acoustical quality of the speech files
- label files (mnemonics and values used)
- lexicon (format, phoneme symbols, completeness)
- speaker information
- environmental information
- transcription quality
- train/testset partitioning

A more detailed account of validation contents follows.

3. VALIDATION OPTIONS

The validation of the first 150 speakers could be limited to the most important items for validation in order to save validation time and costs.
These items are in declining order of importance:

1 completeness of recordings (structurally and incidentally missing files)
2 transcription quality
3 lexicon (format, phoneme symbols, completeness)
4 acoustical quality of the speech files
5 database format/structure and file names
6 label files (mnemonics and values used)
7 speaker information
8 environmental information
9 documentation files
10 train/testset partitioning

The first five items are considered as absolutely necessary by SPEX for a minimal quality guarantee to users who want to use a SpeechDat-Car database for training and testing of speech recognisers.

The correct database structure and filenames are assumed (item 5), otherwise our software cannot proceed to perform the other more important validation tasks.
As a consequence, a database will be immediately resent to the producer when due to such errors the high priority validation tasks cannot be carried out in a straight-forward manner.
The database structure and file names should be clear to the producer from the specifications and from the prevalidation result preceding the first formal validation.

It is recommended that at least the first 5 items are included in a validation.
But also a menu on request (a la carte as it were) is possible.

3. VALIDATION COSTS

In the technical annex a two-step validation procedure (first 150 speakers, next all 300 speakers) was not foreseen, but only the final validation.

In principle a database is checked as whole and cannot be checked in two parts. A database is an integrated package. The end product covers an integrated file structure, lexicon, documentation, speaker and other tables for the whole database. This means:

1. it is not possible just to deliver two separate parts. When the validation centre validates only the second part at the end of the project, it will still not know how the complete database will be.
2. The validation centre cannot split the validation into two equivalent and independent parts and split costs. Each validation is a validation of a complete database, even if it is just half of the complete database. The de-facto difference is only a matter of CPU time for the computer.

A full validation scenario amounts to 6 kECU including prevalidation. Prevalidation alone costs 1 kECU.

This is therefore the price that new partners (new PRLs/owners) should pay for a validation cycle that includes prevalidation and final validation.

For all PRLs (existing and new ones) an extra validation of the first 150 speakers introduces a new validation at corresponding costs.

Costs are calculated according to the table below:

0 data transport from disk and other preparation	0.3 kECU
1 completeness of recordings	1.0
2 transcription quality	1.0
3 lexicon (format, phoneme symbols, completeness)	0.5
4 acoustical quality of the speech files	0.5
5 database format/structure and file names	0.2
6 label files (mnemonics and values used)	0.4
7 speaker information	0.2
8 environmental information	0.5
9 documentation files	0.2
10 train/testset partitioning	0.2

0 should always be payed.

4. REVALIDATION

In case a database is not approved by the Steering Committee after a validation (final validation or validation of first 150 speakers) an extra validation is needed at the costs of the producing partner/owner. This may boil down to a full re-validation or to the re-validation of parts of the database.

The costs of such a revalidation depend on the items to be revalidated, and can thus be directly be calculated from the cost scheme above. As a consequence, the price for a full revalidation is 5 kECU.