SPEECON Deliverable 41

Project ref. no.	*IST-1999-10003*
Project acronym	**SPEECON**
Project full title	Speech Driven Interfaces for Consumer Applications

Security (distribution level)	*Public*
Contractual date of delivery	*M09 = October 2000 (formerly M06)*
Actual date of delivery	*M14 = March 2001*
Deliverable number	*D41*
Deliverable name	*Definition of Validation Criteria*
Type	*Technical Report*
Status & version	*Version 2.0 - (Post-final)*
Number of pages	*24*
WP contributing to the deliverable	*WP4*
WP / Task / Deliverable responsible	*WP4:NSC / D41:NSC*
Other contributors	*Gaël Richard (L&H France)*
Author(s)	*Henk van den Heuvel (SPEX)* *Shaunie Shammass (NSC)* *Ami Moyal (NSC)* *Oren Gedge (NSC)*
EC Project Officer	*Domenico Perrota*

Project Coordinator	Name: *Harald Höge* Company: *Siemens AG, ZT IK 5* Address: *Otto-Hahn-Ring 6, D-81739 München, Germany* Phone: *+49-89-636-44195* Fax: *+49-89-636-49802* E-mail: *harald.h.hoege@mchp.siemens.de* Project web site: *http://www.speecon.com*
Keywords	*Database validation criteria*
Abstract (for dissemination)	This report presents an account of the criteria that databases produced in the framework of SPEECON should meet in order to be accepted as valid and equivalent databases.

DOCUMENT EVOLUTION

Version	Date	Status	Notes
1.0	15/03/00	First draft	Written from Document D1.3.1 of the project LE-8334 (SpeechDat-Car) written by Henk Van den Heuvel (SPEX)
1.1	16/05/00	Draft version, first revision	After comments from Ulm meeting. Addition of a manual check of the Lexicon.
1.2	25/09/00	Draft version, second revision	Comments of Henk van den Heuvel and Shaunie Shammass added, immediately before and after the Paris meeting in Sep. 2000
1.3	5/11/00	Draft version, third revision	Comments of Henk, Shaunie and Ami. Version mailed before Barcelona meeting
1.4	17/11/00	Fourth draft	Results of Barcelona meeting included, together with remarks by A. Kiessling
1.5	12/01/01	Fifth draft	Release for workshop in Warsaw (18-19 Jan.01)
1.6	02/03/2001	Sixth draft	Pre-release for workshop in Tel-Aviv
1.7	12/03/2001	Pre-final	Release for workshop in Tel-Aviv (March 2001)
1.8	20/03/2001	Pre-Final	Corrections after Tel Aviv meeting included
1.9	17/04/2001	Final	Approved after final round
2.0	25/04/2002	*Post-final*	Updated after prevalidation

Table of Contents

1 Executive Summary

This report presents an account of the criteria that databases produced in the framework of SPEECON 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.

The validation procedures and protocols are also described in this report.

2 INTRODUCTION

The objective of this document is to make explicit the criteria that SPEECON 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) and SpeechDat-Car) databases and discussions amongst SPEECON partners. The validation criteria developed within the SpeechDat(II) and SpeechDat-Car projects were used as starting points [7,8].

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. 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 deliverable concerned; the most relevant are D2.1 [1-6] on database specifications. Validation criteria in [1-6] prevail over those in the present deliverable in case of conflicts.

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.

3 DOCUMENTATION

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

- contact person: name, address, affiliation;
- distribution media
  - number of media;
  - contents of each medium;

- layout of the media file system;

- formats of the speech files and of the label files;
- file nomenclature and directory structure;
- reference to the validation report VALREP.DOC;
- speaker recruitment strategies employed;
- prompting
  - presentation design (e.g., which items were spread over a recording session to prevent list effects);
  - prompting example for one recording session;
- database design
  - number of items in the prompting material;
  - a specification of the individual items of the prompting material including lists of vocabulary for each section;
  - for spontaneous items the texts prompted to the speakers should be included together with an English translation;
  - a list of digit forms in the language;
  - connection of prompt items to item numbers in the database (to be provided in the subtitles of individual corpus items);
- recording platform description
  - microphone positions;
  - microphone types;
  - was a high pass filter (ME64 or similar) used or not;
- description of the different recording environments and their distribution in the database
  - indoor environments (office, entertainment);
  - outdoor environments;
  - car environments;
  - children environments;
  - codes used for subdivisions of each environment;
- speaker demographics information
  - which accent 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;

- annotation information
  - procedure used;
  - quality assurance;
  - a list of non-standard and alternative spellings (or reference to file SPELLALT.DOC);
  - standard character set used for transcription (ISO-8859-<n> or other if needed for exotic languages);
  - any other language-dependent information such as abbreviations, proper name conventions, contractions (July or july, isn't, cannot or can not, etc.);
  - annotation symbols for non-speech acoustic events including the standard defined (i.e. [fil], [spk], [sta], [int]) and other language-specific symbols;
  - markers for mispronunciations, recording truncations, unintelligible speech;
- lexicon information
  - procedures to obtain phonemic forms from orthographic input;
  - list of SAMPA phone symbols;
  - list of PinYins and Hepburn Romaji syllables (if applicable);
  - whether or not the transcription and the lexicon are case sensitive;
  - information captured in the phone transcriptions (assimilation and reduction rules);
  - whether multiple transcriptions are supported;
  - if stress information is supplied;
  - if there are any tags, and if so, the tagging conventions used, e.g., record (noun) vs. record (verb);
  - list of words that are from a foreign language;
  - analysis of frequency of occurrence of the phonemes represented in the phonetically rich sentences, phonetically rich words and in the full database (at transcription level); optional for statistics of diphones, triphones; separate tables must be provided for the adults' part of the corpus and the children's part of the corpus;
  - list of rare phonemes;
  - 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 characterize the database.

A template file with section headers and directives of information to be put into each (sub)section will be distributed among partners by the validation centre (SPEX).

4 DATABASE STRUCTURE, FILE NAMES, AND CONTENTS

4.1 File names for label files and speech files and directory names

The database should have the directory structure and file names, as specified in [4], section 4.2, and more specifically in subsections 4.2.4 and 4.2.6.

4.2 The DOC directory

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

The validation of the DESIGN.DOC main documentation file is described in section 3. PLATFORM.DOC contains the most recent platform 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}.TXT contains an overview of all items included for each session per channel. If SUMMAR{1,2,3}.TXT are identical to SUMMAR0.TXT, then they can be omitted.
Also the noise recordings and the silent word recordings should be included in the SUMMAR{0|1|2|3}.TXT files.

The file, VALREP.DOC, containing the validation report is created by the validation centre.

4.3 The TABLE directory

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

The validation of LEXICON.TBL is dealt with in section 8; the validation criteria for the SPEAKER.TBL is given in section 9, and for SESSION.TBL and REC_COND.TBL files in section 10.

4.4 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).

The contents of the README.TXT file differ per database, depending on whether it is for the adults' or the children's database.

Index files should be in \<database_name>\INDEX. The one obligatory file is CONTENT0.LST. CONTENT0.LST should have the correct format, specified in [4], and contain the correct information for the close talk microphone channel.

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

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

4.5 Other requirements

All text files should have <CR><LF> at line ends. This concerns label files, table (.TBL) files, index (.LST) files, and text (.TXT) files.

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

For each label file there should be four corresponding speech files, and for each speech file there should be one corresponding label file.

Obviously, a database should not be infected by any viruses.

5 DATABASE ITEMS AND COMPLETENESS

5.1 Mandatory items specifications

It will be checked if all mandatory items are recorded. The mandatory database items are listed in the tables in [3] and [4]. Each individual item should meet the specifications in [3].

Checks will be directed towards the following at the *prompt* level:

Item type (corpus)	Adults	Children
Application words (101-995/001-214 ; Y01-Y99)	Set size: 450-500 words	Set size: 122 words Each child speaks all words
City / Street names (CO1-2)	275 city names; 275 street names	Each name does not appear more than once
Names (CP1)	150 names	Each name does not appear more than once
E-mail addresses (CW2)	550 addresses	Each name does not appear more than once
Web addresses (CW1)	150 addresses	Each name does not appear more than once
Analogue time phrases (CT1)	Max. 20 specific time words	Max. 20 specific time words
Relative & general dates (CD2)	Max. 50 phrases	Max. 50 phrases
Keyboard characters (CK1-2)	Max. 20 chars (12 fixed for all)	Max. 20 chars (12 fixed for all)
Phon. rich sentences (S01-30/60)	- At least 3300 different sentences - Max. 5 occurrences/sentence	- At least 600 different sentences - Max. 5 occurrences/ sentence
Phon. rich words (W01-W05)	- At least 300 different words - Max. 10 occurrences/word	-
Free spontaneous speech (F01-F30)	At least 10 items on average per session should be present. If not, then at least 2 minutes per session (duration of speech files,	-

	from END label) is required	

- The exact set sizes for the application words per domain are given in section 3.3.5 of [3].

- Formats and ranges of connected digit strings and numbers:
    - Money amounts in local currency (Euro for EC languages), and optionally one or two extra currencies;
        - The main word for the local currency should appear in at least 50% of the prompts of the money amounts (this word is EURO for EC languages). This is a criterion for adult speakers only.

    - All read numeric items should be prompted as words. This holds for items: CI1-4, CB1, CC1-4, CM1, CN1-3, CD1,3, CT1-2; the only exception are the telephone numbers (CE1)

5.2 Validation of missing items

For the *adults'* part of a database it will be checked if all mandatory items are present in sufficient quantities.

- A maximum of 5% of the files of each mandatory item (corpus code) may be *effectively missing*;
- A maximum of 7% of the files of each mandatory item (corpus code) may be *effectively missing* or contain *corrupted* speech only;
- A maximum of 10% of the files of mandatory isolated word items may show a *mismatch* between prompt and transcription text; this percentage includes the effectively missing and corrupted files.
- These criteria are applied only to the annotated close-talk channel.

Effectively missing files are: absent files, and files containing only non-speech (i.e., noise symbols between square brackets and/or items marked as being unintelligible) according to the transcriptions. Files with only *corrupted speech* are files for which *each* word is mispronounced or truncated according to the transcription.

For isolated word items (especially application words, which are 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 (no speech at all or only another or other word(s) instead), then this should be considered as a mismatch. A maximum of 10% of the files may be *mismatching* in this way. It is obvious that effectively missing and corrupted files contribute to this count as well. If the word is present but is transcribed as mispronounced, cut-off or unintelligible, then it is *not* considered as a mismatch.
The following corpus items are involved in this mismatch check:
- application words (adults: 1..-9.., Y.., children: 0..-2.., Y..)
- isolated digit (CI.)

- city name (CO1)
- street name (CO2)
- person name (CP1)
- yes/no items (CQ.)
- keyboard characters (CK.)
- phonetically rich words (adults: W.).

A count of isolated word items that are mispronounced, truncated or unintelligible will be done in order 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 also be carried out (see section 11).

Similar items may compensate for each other in order to meet the completeness criteria. Items with the same corpus identifier AND item identifier (see section 4.2.7 of [4]) can compensate for each other. Exceptions are:

- yes/no questions
- different types of names
- CI1-4, CB1, and CC1-4 can compensate for each other although they do not have the same corpus identifier AND item identifier.

For the *children*'s part of the corpus the same criteria apply, except for the *general words/phrases* (GW) part. For this part the criteria are:

- A maximum of 40% of the files of each mandatory GW item (corpus code) may be *effectively missing*;
- A maximum of 40% of the files of each mandatory GW item (corpus code) may be *effectively missing* or contain *corrupted* speech only;
- A maximum of 40% of the files of mandatory isolated word GW items may show a *mismatch* between prompt and transcription text; this percentage includes the effectively missing and corrupted files.
- These criteria are applied only to the annotated close-talk channel.

5.3 Validation of missing words/digits

The check on the completeness of each corpus code described above in section 5.2 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 or mispronounced. Only if the word is not present in the transcription, it is considered as missing.

As a general rule it is stated that at least 85% of the maximum achievable word tokens should have been recorded.

The following checks will be carried out for the *adults*' part of the corpus:

Corpus item	**Min. Samples required**	**Max samples achievable**
Isolated digits (CI1-4)	0.85*4*550/D[1] per digit = 1870/D per digit	4*550/D[1] per digit
Isolated digit string (CB1)	0.85*550 per digit (=0.85*D*550/D) =467 per digit	550 per digit
Connected digit strings (CC1-4)	0.85*(4*5)*550/D per digit =9350/D per digit	(4*5)*550/D per digit
Application words (101-995, Y01-Y99)	190 per word	220 per word
Spelt letters (CL1-3)	0.5 * (3*7*550) /L[2] per letter =5775/L per letter	
Phon. rich sentences (S01-S30)	500 tokens/phone	
Phon. rich words (W01-05)		
Dates (analog and digital formats) (CD1, CD3)	35 per month name 60 per day name	48 per month name 78 per week day
Yes/No answers (CQ1-2)	465 yes / 465 no	550

[1]D being the number of digits for a language, and D may differ per category; [2]L being the number of letters for a language.

For the combined phonetically rich words and phonetically rich sentences the following criteria apply for adult speakers:

- Each phoneme is spoken in at least 90% of the 550 sessions. Exception: rare phonemes:
  - these appear mainly in loan words AND
  - a max. of 10% of all phonemes in the language may be rare

Geminates (duration oppositions, like /m/ and /mm/ in Italian), and vowel composita (quasi diphthongs, like /a:6/ in German and /iu/ in Dutch) are excluded from this count, since they can be synthesized from other phonemes in the language.

For the *children* the following criteria apply:

Item type (corpus)	Min. Samples required	Max samples achievable
Isolated digits (CI1-4)	0.60*4*50/D[1] per digit = 120/D per digit	4*50/D[1] per digit
Isolated digit string (CB1)	0.60*50 per digit (=0.60*D*50/D) = 30 per digit	50 per digit
Connected digit strings (CC1-4)	0.60*(4*5)*50/D per digit = 600/D per digit	(4*5)*50/D per digit
Application words (001-214; Y01-Y99)	40 per word	50 per word
Spelt letters (CL1-3)	0.5 * (3*7*50) /L[2] per letter = 525/L per letter	
Phon. Rich sentences (S01-60)	50 tokens/phone	
Dates (analog and digital formats) (CD1, CD3)	-	-
Yes/No answers (CQ1-2)	45 yes / 45 no	50

[1]D being the number of digits for a language, and D may differ per category; [2]L being the number of letters for a language.

- Each phoneme is spoken at least 50 times. Exception: rare phonemes:
  - these appear mainly in loan words AND
  - a max. of 10% of all phonemes in the language may be rare

Geminates (duration oppositions, like /m/ and /mm/ in Italian), and vowel composita (quasi diphthongs, like /a:6/ in German and /iu/ in Dutch) are excluded from this count, since they can be synthesized from other phonemes in the language.

6 ACOUSTIC QUALITY OF THE SPEECH FILES

SNR evaluation (SNR value as measured by Sony software during the recording):

1. At least 90% of the files of all sessions should have an SNR of 15 dB or more for the *close talk* channel
2. If the 80 Hz high-pass preamplifier filter is switched off for the car recordings then at least 80% of the files of all sessions recorded in cars should have an SNR of 10 dB or more for the *close talk* channel; If the HP filter is switched on, then 1 above is valid.
3. The children's environment is not checked for SNR level
4. The filter should be switched on or off for the WHOLE database, and this should be documented in the DESIGN.DOC
5. For recordings made before October 2001, a list of sessions with filter on/off should be provided in DESIGN.DOC, if the filter switch was variable over sessions until then.

In addition, the noise level value measured during recording and stored in the label files is used for evaluation:

- at least 90% of the sessions recorded in the *Office* environment must have a noise range between 30-60 dB(A)
- at least 90% of the sessions recorded in the *Entertainment* environment must have a noise range between 30-65 dB(A)
- at least 90% of the sessions recorded in the *Children's* environment must have a noise range between 30-70 dB(A)
- at least 90% of the sessions recorded in the *Public Place* environment must have a noise range between 45-90 dB(A)
- at least 90% of the sessions recorded in the *Car* environment must have a noise range between 28-80 dB(A)

Saturation level will be handled by the collection software.

Each file should contain a leading and a trailing silence portion. This will be tested automatically or manually on the 2000 files selected for transcription validation (see section 11). The validation criteria are:

- at least 500 ms before and after the speech portion
- check on close-talk microphone only
- at least 90% of the tested files should satisfy the criterion.

A final requirement is that:

- A set of noise recordings should be present (_01- _06) for each new (sub)environment (recording position).
- For (new recording positions in) PUBLIC places only _01- _03 should be present.
- For CAR environments there should be no noise recordings.

7 ANNOTATION FILES

Checks will be performed as to:

- Correct use of labels 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], section 4.2.10.

Special requirements for the DBA label:

- For the silent word recordings, DBA should have correct/appropriate values;
- For the noise recordings (impulse response files), DBA is not validated and may be nil.

8 LEXICON

8.1 Format checks

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 distorted words
- If tagging is supplied, check that all tag symbols are defined and only those symbols are used.

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

The lexicon should be complete. The completeness check is carried out on the orthographic transcriptions in the label files in order to find out if all transcribed words are in the lexicon. Undercompleteness is not permitted, overcompleteness is.

8.2 Validation of phonemic transcriptions

1000 lexicon entries should be checked for phonetic correctness by native speaker phoneticians that were not involved in the original transcription process, or by comparing with other available pronunciation lexicons.

The validation of the phonemic correctness of the lexicon entries is organised as follows:

- 1000 entries are randomly extracted from the lexicon;
- Of phonemic transcriptions only the first one is kept;

- The check is carried out at the segmental level only (not on syllable boundaries or stress marks, if provided);
- The check is carried out by a phonetically educated person who is a native speaker of the language;
- The given transcription receives the benefit of the doubt;
- The given transcription is correct if it represents a possible pronunciation of the word (which is not necessarily the most common);
- Each transcription is rated on a 3-point scale: OK; Minor error; Severe error;
- A max. of 10% minor errors is allowed; and a max. of 5% severe errors is allowed;
- A minor error occurs if only one symbol in the transcription is wrong;
- A severe error occurs if more than one symbol is wrong.

Since only a sample of 1000 entries is evaluated, the detected errors give the following confidence intervals when extrapolated to the entire DB.

Error percentage	**Confidence interval**
5%	3.6% - 6.4%
10%	8.1% – 11.9%
50%	46.9% - 53.1%
95%	93.6% - 96.4%

9 SPEAKER INFORMATION AND DISTRIBUTION

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

A minimum number of speakers of 600 should be recorded: 550 adults and 50 children.

Below we summarise the speaker distribution criteria as given in [5].

9.1 Speaker sex

- For the adults the sex misbalance may be 5% at maximum for the total database of 550 speakers and for each *age* category. This means that the proportion of sessions from male and female speakers must be in the interval 45-55% for both sexes for each age category; the allowed interval for the whole adults' database is thus [248-303] speakers per sex.

- For each *dialectal region* the proportion of each sex should be between 30% and 70% for each region.

- For each *recording environment* the proportion of each sex should be between 30% and 70% for each environment.

For children gender is not considered relevant.

9.2 Speaker age

For adult speakers the following criteria are valid:

Age interval	Proportion of speakers	Requirement
15-30	≥ 30%	45-55% male
31-45	≥ 30%	45-55% male
46+	≥ 10%	45-55% male

Boys of 15+ should have had the voice break.

For children the following criteria apply:

Age interval	Proportion	Requirement
08-10	≥ 30%	Mandatory
11-15	≥ 30%	Mandatory No boys with voice breaks may be included

9.3 Speaker accent distribution

Dialect requirements concern only adult speakers.

- A database contains between 4 and 6 accent regions or dialects.
- Each accent region is represented by at least 0.70 * 550/D speakers, D being the number of dialects distinguished in the database
- In the office and public place environments (200 speakers each), each dialect region is represented by at least 0.5 * 200 / (number of dialects) speakers.

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

10 RECORDING CONDITIONS

The session table and the recording condition table files should have the format specified in [4].

Environments:

	Environment	#Sessions
Home	Office	200 [190-210]
	Entertainment	75 [71-79]
	Children	50 [47-53]
Mobile	Car	75 [71-79]
	Public Places	200 [190-210]

For each environment a deviation of max. plus or minus 5% is allowed as indicated in the table above. The total number of sessions should, of course, remain 600 (or more).

Each environment is divided into sub-environments. The criteria for the features of each sub-environment and the permitted division of speakers over the sub-environments is given in the tables presented in section 2.4.2 of [2].

11 TRANSCRIPTION

11.1 Type of errors

Three types of errors are distinguished:

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

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 transcription errors.

The transcription validation of speech is carried out by a neutral trained native speaker of the language concerned who did not participate in the original transcription process. The transcription validation of non-speech symbols is not necessarily done by a native speaker of the language, but by someone experienced in listening to background noises and capable to decide which noises should be transcribed or not. 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.

It is the first (native) validator who also checks for channel mismatches. One file of the other channels will be linked to each tested file of the close-talk channel. A channel mismatch means that recordings that are supposedly simultaneous do not contain the same utterance or contain only part of the same utterance.

11.2 Transliterations

The following criteria are valid for 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, unless there is drop-out).

The full description of the relevant transcription conventions can be found in [4].

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

11.3 Criteria for validation

The main criteria for the validation of the transcriptions by the expert 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;
- A maximum of 5% channel mismatches may be found.

As errors in non-speech we consider both erroneous omissions and insertions of noise symbols.
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.
The error percentage is only determined on item level, not on word level.

11.4 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. 20% of the long items must be from the spontaneous speech.
The following corpus items are considered as short items: single word utterances (application words, single digits, Y/N questions, names, phonetically rich words). All other items are considered as long items.

For each set of 1000 items the (95%) confidence intervals for varying error percentages are:

Error percentage	**Confidence interval**
5%	3.6% - 6.4%
10%	8.1% – 11.9%
50%	46.9% - 53.1%
95%	93.6% - 96.4%

And for the full set of 2000 items the confidence intervals are:

Error percentage	Confidence interval
5%	4.0% - 6.0%
10%	8.7% - 11.3%
50%	47.8% - 52.2%
95%	94.0% - 96.0%

11.5 Spelling check

A formal spelling check of the orthographic transcriptions 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.

12 Validation procedures

A database is validated in at least three stages: prevalidation, validation and pre-release validation.

12.1 Prevalidation

The delivery for prevalidation contains two parts:

A. The prompt files as envisaged for the full set of 600 sessions, with a clear distinction of adults' and children's material. Also the lexicon table file for all read items should be included.
B. A complete minidatabase of 10 sessions, 2 sessions per environment (incl. 2 children). 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 goals of the prevalidation 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;

The following checks are carried out for packages A and B, respectively:

A.	B.
The completeness checks as described in section 5, as far as possible for read material	All checks described in sections 3-10 as far as possible, typically not including completeness checks for corpus items, speaker and rec. environment distributions
The lexicon checks as described in sections 8.1 *and* 8.2	The automatic check on transcription symbols

	Quick check on the use of non-speech transcription symbols by a non-native speaker

12.2 Validation

For validation the procedure is as follows:

1. The producing partner sends a CD-ROM with all files, except for speech files, to the validation centre (also the label files must be included!).
2. Immediately after receiving 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 a CD and also sent to SPEX. Alternatively, the producing partner may choose to send the CONTENT0.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. Immediately after receiving the database, a quick check verifies if all needed files are present and if the label files have the correct format.
4. After a successful quick check, all checks described in the previous sections 3-10 are carried out when the database gets out of the queue.
5. The partner may be requested to send some additional sessions for acoustic quality evaluation (section 6).
6. The result of the validation, the validation report VALREP.DOC, is sent to the producing partner.
7. 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.

12.3 Revalidation

In case a database is not approved or only conditionally approved by the Consortium then an extra validation is needed at the costs of the producing partner/owner.
This may boil down to a full revalidation or to the revalidation of parts of the database.

The costs of such a revalidation depend on the items to be revalidated. A table for partial revalidation is shown below:

0 data transport from disk and other preparation	0.2 kEuro
1 completeness of recordings	1.0
2 transcription quality (orthographic)	0.8
3 transcription quality (phonemic)	0.8
4 lexicon (format, phoneme symbols, completeness)	0.2
5 acoustical quality of the speech files	0.5
6 database format/structure and file names	0.2
7 label files (mnemonics and values used)	0.4

8 speaker information	0.2
9 environmental information	0.5
10 documentation files	0.2

As a consequence, the price for a full revalidation is 5 kEuro.

If the revalidation work is minor and can be completed in -say- less than one 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 invoice.

Full validations of other databases always have priority over revalidations in our planning. Revalidations will be put at the back of the queue, and only be scheduled before a normal full validation if there is an empty time slot in which there is nothing else to validate. A revalidation, once started, should not last longer than 3 weeks.

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.

12.4 Pre-release validation

When a database is approved, the final masters must be made. Prior to multiplication, SPEX will carry out one additional check on the non-speech data disk. This validation includes the following checks:

- Structure of CD;
- Version of DESIGN.DOC;
- Version of all re-submitted files;
- Version of the validation report.

Once this disk is approved, multiplication and distribution of the database can commence.

13 REFERENCES

[1] Andreas Kiessling, Frank Diehl, Volker Fischer, Krzysztof Marasek*: Specification of databases – Introduction.* SPEECON Technical Report D2.1.1.

[2] Frank Diehl, Volker Fischer, Andreas Kiessling, Krzysztof Marasek*: Specification of databases – Specification of recording scenarios.* SPEECON Technical Report D2.1.2.

[3] Krzysztof Marasek, Frank Diehl, Volker Fischer, Andreas Kiessling*: Specification of databases – Specification of corpus and vocabulary.* SPEECON Technical Report D2.1.3.

[4] Volker Fischer, Frank Diehl, Andreas Kiessling, Krzysztof Marasek*: Specification of databases – Specification of annotation.* SPEECON Technical Report D2.1.4.

[5] Andreas Kiessling, Frank Diehl, Volker Fischer, Krzysztof Marasek*: Specification of databases – Specification of speakers.* SPEECON Technical Report D2.1.5.

[6] Andreas Kiessling, Frank Diehl, Volker Fischer, Krzysztof Marasek*: Specification of databases – Specification of language specific items.* SPEECON Technical Report D2.1.6.
[7] Henk van den Heuvel*: Validation criteria*. SpeechDat(II) Technical Report SD1.3.3. Version 1.9, 1997.
[8] Henk van den Heuvel*: Validation criteria*. SpeechDat-Car Technical Report D1.3.1. Version 3.4, 1999.