SPEX / Dept. of Language and Speech
University of Nijmegen
Erasmusplein 1
NL-6525 HT Nijmegen
The Netherlands
e-mail: spex@spex.nl

SUBJECT:	Validation Czech FDB SpeechDat(E) corpus
AUTHORS:	Henk van den Heuvel, Mieke van Wijck
VERSION:	1.2
DATE:	19 October 2000

INTRODUCTION

The speech databases made within the SpeechDat(E) project were validated by SPEX, Nijmegen, the Netherlands, to assess their compliance with the SpeechDat format and content specifications, as documented in Deliverables 1.12, 1.3 and 1.4.2 of the project.

The validation results of the Czech Fixed Network SpeechDat(E) database (1000 speakers) are contained in this document.

This database was validated and approved by the SpeechDat(E) Consortium.

In the validation procedure we systematically check a list of validation criteria for a range of subjects. In the following sections we will evaluate these criteria one by one. Validation results that call for attention because of deviations from the SpeechDat specifications are marked by
=>
so that they can be easily found.

The following subjects were validated:

The document is concluded by

1 DOCUMENTATION

- File DESIGN.DOC; & deliverables ED1.12.?, ED1.3 and ED1.4.1 can be handy
  OK ED1.3 is also included

- Appendix with language specific information may be in PDF or PS format
  OK An Annex file which contains all prompted material is provided in both PS and PDF format.

- Language of doc file: English
  OK

- Contact person: name, address, affiliation
  OK

- Number of CD's
  OK section 1

- Contents of each CD
  OK, section 1 of DESIGN.DOC

- The directory structure of the CD's
  - database, block and session orderings
    OK, section 1.3
  - directories DOC, INDEX, TABLE (and optionally PROMPT, SOURCE)
    OK, section 1.3

- The format of the speech files (A-law, 8 bit, 8 kHz, uncompressed)
  OK, section 1

- File nomenclature
  - root files
    OK, section 1.3
  - names of speech files and label files
    OK, section 1.2
  - files in directories DOC, INDEX, TABLE (and optionally PROMPT, SOURCE)
    OK, section 1.3

- Contents and format of the label files
  - Clarification of attributes (three letter mnemonics)
    OK, section 1.4.
  - example of label file
    OK, section 1.4

- Recording platform
  OK, section 2.1

- Speaker recruitment
  OK, section 2.2

- Prompting
  - connection of sheet items to item numbers on CD
    OK, section1.2.
  - sheet example
    OK, section 10
  - items must be spread over the sheet to prevent list effects (strong recommendation)
    OK

- Description of all recorded items
  OK, section 3

- Analysis of frequency of occurrence of the sub-word units (phones) represented in
  - the phonetically rich sentences;
    OK, section 3.12
  - the phon. rich words;
    OK, section 3.13
  - the full database;
    OK, section 3.13

  the last count should be made on all items in the database on transcription level. A listing and motivation of rare phones should be given and, if possible, a mapping of these rare phones onto more frequent ones
  OK

- Transcription conventions
  - procedure
    OK, section 4
  - quality assurance
    OK, section 4
  - character set used for annotation (transcription) (ISO-8859-?)
    OK, section 4
  - annotations symbols for non-speech acoustic events must be mentioned at least for Filled Pause, Speaker Noise, Stationary Noise, Intermittent Noise
    OK, section 4
  - list of symbols used to denote word truncations, mispronunciations and not understandable speech
    OK, section 4
  - case sensitivity of transcriptions
    OK, section 4
  - use of punctuation
    OK, section 4

- Lexicon information
  - Procedures to obtain phonemic forms from orthographic input (lexicon generation and lay out)
    OK, section 6
  - splitting of entries only at spaces
    OK, section 6
  - (Reference to) SAMPA symbols used
    OK, section 6
  - case sensitivity of entries (matching the transcriptions)
    OK, section 4, but explicitly in section 6

- Speaker demographics
  - which regions, how many of each
    OK, section 5.1
  - motivation for selection of regions
    OK, section 5.1
  - which age groups, how many of each
    OK, section 5.2
  - sexes: males, females, also children?; how many of each.
    OK, section 5.2
  - how many sessions by how many speakers
    OK, uniqueness of speakers is assumed.

- Recording conditions
  OK, section 7

- Information on test (set) specification
  OK, section 8

- The validation report made by SPEX (VALREP.{TXT,DOC}) is referred to
  OK, section 1.3

2 DATABASE STRUCTURE, CONTENTS AND FILE NAMES

– Directory / subdirectory conventions
  Format of directory tree should be
  \<database>\<block>\<session>
  - database: defined as <name><#><language code><name> is FIXED
    <#> is 3 for SpeechDat(E)
    <language_code> is the ISO two-letter code for the language
    OK
  - block: defined as BLOCK<nn> where <nn> is a progressive number from 00 to 99.
    Block numbers are unique over all CD's.
    They correspond to the first two digits of <nnnn> below.
    OK
  - session: defined as SES<nnnn> where <nnnn> is the session code also appearing in file name
    OK

– File naming conventions
  All file names should obey the following pattern: DDNNNNCC.LLF
  - DD: database identification code
    For SpeechDat(E): A3 = fixed net
    OK
  - NNNN: session code 0000 to 9999
    OK
  - CC: item code; first character is item type identifier, second character is item number
    OK
  - LL: ISO-639 language code (with extensions)
    OK
  - F: speech file type
    A is for A-law
    O is for Orthographic label file
    OK

– NNNN in filenames is not in conflict with BLOCK and SES numbers in pathname
  OK

- Correct item codes should be used:
  A1-6 : common application words
  B1 : sequence of isolated digits
  C1 : prompt sheet number
  C2 : telephone number
  C3 : credit card number
  C4 : PIN code
  D1-3 : dates
  E1 : application word phrase
  I1 : isolated digit
  L1-3 : spelled words
  M1-2 : money amounts
  N1 : natural number
  O1 : spontaneous name
  O2 : city of call/birth
  O3 : most frequent city name
  O5 : most frequent company/ agency name
  O7 : forename & surname
  O8 : surname
  Q1-2 : yes/ no questions
  S1-9 : phonetically rich sentences
  Z0-1 : idem
  T1 : time of day
  T2 : time phrase
  W1-4 : phonetically rich words
  OK

- Contents lowest level subdirectories should be of one call only
  OK

- All text files should be in MS-DOS format (<CR><LF>) at line ends
  OK

- A README.TXT file should be in the root describing all (documentation) files on the CD-ROM
  OK

- 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. Example of contents: FIXED1EN_01
  OK

- A copyright statement should be present in the file COPYRIGH.TXT (root)
  OK

- Documentation should be in \<database_name>\DOC
  - DESIGN.DOC
    OK
  - TRANSCRP.DOC (optional)
    Not provided
  - SPELLALT.DOC (optional)
    Not provided
  - SAMPALEX.PS
    OK
  - ISO8859<nr>.PS
    OK, ISO88592.PS is present
  - SUMMARY.TXT
    OK
  - SAMPSTAT.TXT
    OK

- The contents list (CONTENTS.LST) is in \<database_name>\INDEX
  OK

- Tables should be in \<database_name>\TABLE
  - SPEAKER.TBL
    OK
  - LEXICON.TBL
    OK
  - REC_COND.TBL (optional)
    Not provided
  - SESSION.TBL (optional)
    OK

- Index files (optional) should be in \<database_name>\INDEX
  Mandatory are:
  - CONTENTS.LST
    OK
  - A3TST<language code>.SES
    OK

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

- All sessions indicated in the documentation SUMMARY.TXT are present on the CDs
  OK

- Contents lowest level subdirectories should be of one call only
  OK

- Empty (i.e. zero-length) files are not permitted
  OK

- File match: For each label file there must be one speech file and vice versa
  OK

- Part of the corpus is designed for training and a smaller part for testing
  OK, A3TSTCS.SES defines 200 test sessions, and A3TRNCS.SES
  852 train sessions. Both sets do not overlap and the files have the correct format.

- All table files, and index files should report the field names as the first row in the files using tabs as in the data records following.
  OK

- The contents of the database as given in CONTENTS.LST should comprise:
  - CD-ROM volume name (VOL:)
    OK
  - full pathname (DIR:)
    OK
  - speech file name (SRC:)
    OK
  - corpus code (CCD:)
    OK
  - speaker code (SCD:)
    OK
  - speaker sex (SEX:)
    OK
  - speaker age (AGE:)
    OK
  - speaker accent (ACC:)
    OK
  - orthographic transcription of uttered item (LBO:)
    OK
  - The first line should be a header specifying the information in each record.
    OK
  - This file must be supplied as an ASCII TAB delimited file.
    OK

- The contents of the SUMMARY.TXT files should comprise:
  - The full directory name where speech and label files are to be found
    OK
  - the session number
    OK
  - a string of typically N codes. Each item present is represented by its code. If the item is missing, a '--' should appear.
    OK
  - recording date
    OK
  - recording time of first item
    OK
  - optional comment text
    OK
  - all these fields are separated by spaces
    OK

  Note: The contents of the SUMMARY.TXT file are not CD-dependent.
  OK

- Missing items per session
  Check with documentation (SUMMARY.TXT)
  OK

- The database should not contain any viruses
  OK

3 ITEMS

A. Check on mandatory corpus items

– 6 common application words (code A1-6)
  - read
    OK
  - set of 25-30 should be used, 25 of which are fixed for all
    OK, 29 are used
  - for Russian 33 words are used
    Not applicable
  - minimum number of examples of each word = #speakers/10
    (this is a soft target)
    OK, more than 200 occurrences for each word were found.

– 1 isolated digit (code I1)
  - read or prompted
    OK

– 1 sequence of 10 isolated digits (code B1)
  - each sequence must include all digits
    OK
  - optional are hash and star
    Not used

– 4 connected digits (code C1-4)
  – 5+ digit number to identify the prompt sheet
  - read
    OK
  – 9-11 digit telephone number
    - read
      OK
    - local numbers
      OK
    - inclusion of GSM numbers recommended
      OK
  – 16 digit credit card number
    - read
      OK
    - set of 150
      OK, each number appears between 5 and 9 times
    - if there is a checksum then formula must be provided
      Not needed

- 6 digit PIN code
  - read
    OK
  - set of 150
    OK, each number appears between 4 and 9 times.
  - ~30 digits per call are required
    OK
  - digits must appear numerically on the sheet, not as words
    OK

- 1 date (code D1)
  - spontaneous
    OK

- 1 date (code D2)
  - read, wordstyle
    OK
  - analogue form
    OK
  - covering all weekdays and months,
    ordinals and year expressions (also exceeding 2000)
    OK, there is some variation in the occurrences of day names:
    středa: 109; pátek: 206

- 1 relative date (code D3)
  - read
    OK
  - analogue
    OK
  - should include forms such as
    TODAY, TOMORROW, THE DAY AFTER TOMORROW, THE NEXT DAY,
    THE DAY AFTER THAT, NEXT WEEK, GOOD FRIDAY, EASTER MONDAY, etc.
    OK, each occurs between 34 and 40 times

- 1 application word phrase (code E1)
  - application word is embedded in phrase
    OK
  - read or spontaneous
    OK, read
  - a length of min. 3 words per sentence is strongly recommended
    OK

– 3 spelled words (code L1-3)
  - L1 is spontaneous name spelling linked to O1
    (or to another item explicitly documented)
    OK
  - others are read
    OK
  - equal balance of all vocabulary letters
    artificial words can be used to enforce this balance
    OK, each letter appears at least 300 times.
  - average length at least 7 letters
    OK, average length of the (read) spelled items is 8.67
  - may include names, cities and other frequently spelled items
    OK, L2 is a city name
  - should include equivalents of:
    A-Z, accent words, DOUBLE, APOSTROPHE, HYPHEN
    OK

– 2 money amounts (code M1-2)
  - read
    OK
  - currency words should be included
    OK
  - mixture of small amount including decimals
    and large amounts not including decimals
    OK
  - M1 should contain national currency
    OK
  - M2 should contain 50% dollar amounts and 50% Euro amounts
    OK
  - All amounts for prices between a news paper and a car
    Unknown

– 1 natural number (code N1)
  - read
    OK
  - provided as numbers (numerically)
    OK
  - decimal numbers are only allowed for additional natural numbers (strong recommendation)
    Decimals are used in the one natural number per call available
  - max. of 4 significant digits for values more than 1,000,000
    OK

– 6 directory assistance names (code O1-8)
  - 1 spontaneous name (e.g. forename)
    OK
  - 1 spontaneous city name
    OK
  - 1 read city name (list of at least 500 most frequent)
    OK, 499 different names were retrieved in the database.
  - 1 read company/agency name (list of at least 500 most frequent)
    OK, 496 different names were retrieved in the database.
  - 1 read proper name, fore- and surname
    (list of 150 names: both male and female names)
    OK
  - 1 read surnames
    (list of 150 names: both male and female names)
    OK

– 2 yes/ no questions (code Q1-2)
  - spontaneous, not prompted
    OK
  - one question should elicit (predominantly) 'no' answers;
    the other (predominantly) 'yes' answers
    OK
  - also fuzzy answers should be envisaged
    OK

– 12 phonetically rich sentences (code S0-9, Z0-1)
  - read
    OK
  - minimum number of phone examples = #speakers/10
    OK
  - Russian needs only 9 sentences
    Not applicable
  - each sentence may appear a max. of 10 times
    OK
  - minimum number of different sentences: 1200
    (for Russian: 2250)
    OK, 5285 different sentences were found

– 1 time of day (code T1)
  - spontaneous
    OK

– 1 time phrase (code T2)
  - read
    OK
  - analogue form
    OK
  - equal balance of all words
    OK
  - should include equivalents of:
    AM/ PM, HALF/ QUARTER, PAST/ TO, NOON, MIDNIGHT, MORNING, AFTERNOON, EVENING, NIGHT, TODAY, YESTERDAY, TOMORROW
    OK, 'yesterday' and 'tomorrow' have been included in item D3.

– 4 phonetically rich words (code W1-4)
  - read
    OK
  - minimum number of phone examples = #speakers/10
    OK
  - each word may appear a max. of 5 times
    => We found 34 words occurring more than 5 times in the database.
    => The distribution is as follows:
    1 word appears 11 times;
    1 word appears 8 times;
    13 words appear 7 times;
    19 words appear 6 times.

  - minimum number of different words: 800
    (for Russian: 2000)
    OK, 1921 different words were found.

B. Checks on presence of corpus files

The following completeness checks are performed on obligatory SpeechDat items only:

1. Structurally missing corpus items
   OK, none of the mandatory items is missing.

   On the contrary, a number of additional, optional items is recorded:
   X1: school district of speaker
   X2: speaker sex
   X3: district where speaker called from
   X4: type of handset used by speaker

2. Incidentally missing files
a. files that are not there

We found 133 missing files, according to the following distribution over the corpus items:
A1: 1
C3: 1
D1: 1
D3: 1
I1: 2
L1: 19
L2: 14
L3: 7
M1: 6
N1: 8
O1: 1
O2: 2
O3: 1
O5: 4
Q1: 1
S0: 1
S1: 2
S5: 2
S6: 1
T1: 5
T2: 3
W3: 5
W4: 2
X1: 10
X2: 10
X3: 12
X4: 11

b. files with empty transcriptions in the LBO label field (effectively missing files)

We found 2 files with empty transcription (only noise symbols and/or **),

If we merge these files with the missing files (being 0) given above then we get the following distribution:

1 A1
1 C3
1 D1
1 D3
2 I1
19 L1
15 L2
7 L3
6 M1
8 N1
1 O1
3 O2
1 O3
4 O5
1 Q1
1 S0
2 S1
2 S5
1 S6
5 T1
3 T2
5 W3
2 W4
10 X1
10 X2
12 X3
11 X4

c. corrupted speech files

If we regard utterances which have only truncated or mispronounced words as corrupted files, and merge these with the effectively missing files under b. then the following distribution emerges:

19 A1
5 A3
2 A5
1 C3
1 D1
3 D3
4 I1
22 L1
17 L2
7 L3
6 M1

8 N1
4 O1
7 O2
5 O3
7 O5
1 O7
2 O8
1 Q1
3 Q2
1 S0
2 S1
2 S5
1 S6
5 T1
4 T2
10 W1
14 W2
16 W3
7 W4
11 X1
13 X2
13 X3
12 X4

d. files containing truncation and mispronunciation marks

(*, **, ~ are counted in the transcriptions of the individual items to get an idea of distorted speech data. This will not be used to reject or approve a database but it will be supplied as supplementary information.)

We found 1344 transcriptions with at least one *, or **, or ~,
according to the following distribution over the items:

A1: 52
A2: 6
A3: 17
A4: 3
A5: 5
A6: 3
B1: 11
C1: 59
C2: 23
C3: 30
C4: 53
D1: 48
D2: 18
D3: 10
E1: 42

I1: 18
L1: 13
L2: 11
L3: 24
M1: 36
M2: 24
N1: 15
O1: 21
O2: 16
O3: 15
O5: 27
O7: 13
O8: 6
Q1: 21
Q2: 11
S0: 47
S1: 35
S2: 47
S3: 36
S4: 54
S5: 42
S6: 52
S7: 47
S8: 54
S9: 39
T1: 18
T2: 20
W1: 18
W2: 26
W3: 17
W4: 12
X1: 6
X2: 6
X3: 5
X4: 8
Z0: 57
Z1: 47

3. Overall conclusion

SpeechDat has the following criteria for missing items:

- A maximum of 5% of the files of each mandatory item (corpus code) may be effectively missing.
- As missing files are counted: absent files, and files containing non-speech events only.
- For the phon. rich sentences a maximum of 10% of the files may be effectively missing or corrupted
- There will be no further comparison of prompt and transcription text in order to decide if a file is effectively missing.
  As a consequence: If there is some speech in the transcription, then the file will NOT be considered missing, even if it is in fact useless.

Since the Czech database contains 1000 calls, each (obligatory) item may effectively miss a maximum of 50 files.

For the decision of completeness of an item the distribution given in 2b above should be used. Thus it can be concluded that all recorded items are sufficiently complete. This even holds if the corrupted items are included into the counts.

For the completeness of the phonetically rich sentences the distribution given for 2c applies. It can be seen that also this criterion is easily fulfilled.

4 SAMPLED DATA FILES

1. Coding
   - A-law, 8 bit, 8 kHz, no compression

   OK

2. Sample distribution

Several sample statistics are generated: File length, clipping rate, mean sample value, Signal-to-Noise Ratio (SNR). Statistics were generated on file level by the producer of the database, using SPEX software. The results were delivered to SPEX. SPEX compiled histograms on the basis these results. These histograms are presented below, both on file level and on directory (call) level. The histograms are presented as they are and not further interpreted by SPEX. On the basis of these data the user of the database should be able to decide which acoustic quality is still acceptable for the application at hand. Statistics on the acoustics of individual speech files can be retrieved from file \DOC\SAMPSTAT.TXT.

The columns in SAMPSTAT.TXT have the following meaning:

File max min #samples cliprate mean snr
A31001C2.RUA:16384:-13056:80000: 0.00: -4.28: 35.89

2.1 File length

We calculated the length of the files in seconds in order to trace spurious recordings if files were of extraordinary length.

Duration distribution over all items:

Length (s) #Occurrences

2 - 3 : 1152
3 - 4 : 14706
4 - 5 : 8240
5 - 6 : 6592
6 - 7 : 8073
7 - 8 : 5115
8 - 9 : 3334
9 - 10 : 1784
10 - 11 : 1510
11 - 12 : 985
12 - 13 : 820
13 - 14 : 537
14 - 15 : 460
15 - 16 : 323
16 - 17 : 178
17 - 18 : 119
18 - 19 : 109

19 - 20 : 72
20 - 21 : 333
21 - 22 : 15
22 - 23 : 13
23 - 24 : 7
24 - 25 : 5
25 - 26 : 89

Duration distribution over calls/directories:

Length (s) #Occurrences

4 - 5 : 37
5 - 6 : 522
6 - 7 : 327
7 - 8 : 114
8 - 9 : 33
9 - 10 : 15
10 - 11 : 3
11 - 12 : 1

None of the calls has an alarmingly high average file length.

2.2 min-max samples
We provide a histogram with clipping ratios. The clipping ratio is defined as the proportion of samples in a file that is equal to the maximum/ minimum value, divided by all samples in the file.
The histogram, then, is an overview of how many files were found in a set of clipping rate intervals.

Clip distribution for all items:

Clipping rate (in %) Occurrences

0.0 - 0.1 : 4929
0.1 - 0.2 : 1035
0.2 - 0.3 : 542
0.3 - 0.4 : 256
0.4 - 0.5 : 226
0.5 - 0.6 : 142
0.6 - 0.7 : 105
0.7 - 0.8 : 68
0.8 - 0.9 : 78
0.9 - 1.0 : 44
1.0 - 1.1 : 49
1.1 - 1.2 : 37

1.2 - 1.3 :	33
1.3 - 1.4 :	29
1.4 - 1.5 :	17
1.5 - 1.6 :	6
1.6 - 1.7 :	25
1.7 - 1.8 :	13
1.8 - 1.9 :	9
1.9 - 2.0 :	10
2.0 - 2.1 :	3
2.1 - 2.2 :	14
2.2 - 2.3 :	8
2.3 - 2.4 :	2
2.4 - 2.5 :	4
2.5 - 2.6 :	10
2.6 - 2.7 :	6
2.7 - 2.8 :	1
2.8 - 2.9 :	6
2.9 - 3.0 :	1
3.0 - 3.1 :	3
3.2 - 3.3 :	1
3.4 - 3.5 :	4
3.5 - 3.6 :	2
3.6 - 3.7 :	2
3.7 - 3.8 :	1
3.8 - 3.9 :	1
4.3 - 4.4 :	1
4.4 - 4.5 :	1
4.5 - 4.6 :	1
4.6 - 4.7 :	2
5.2 - 5.3 :	1

Number of files with absolute maximum < 32256: 46843

Clip distribution over calls/directories:

Clipping rate (in %)	Occurrences
0.0 - 0.1 :	437
0.1 - 0.2 :	24
0.2 - 0.3 :	10
0.3 - 0.4 :	8
0.4 - 0.5 :	2
0.5 - 0.6 :	3
0.6 - 0.7 :	1
0.7 - 0.8 :	1
0.8 - 0.9 :	2
1.0 - 1.1 :	1

1.1 - 1.2 : 1
1.4 - 1.5 : 3
1.6 - 1.7 : 1

Number of directories with absolute maximum < 32256: 558

There are six calls with an average clipping rate higher than 1.0%:

SES0135: 1.12
SES0397: 1.08
SES0418: 1.67
SES0419: 1.45
SES0826: 1.41
SES0875: 1.42

However, these sessions are of acceptable quality, so we recommend to put them in the test set. Therefore, we recommend to modify the files A3TSTCS.SES and A3TRNCS.SES accordingly.

2.3 Mean values

We computed the mean sample value of each item in each call. We provide a histogram with mean values below. The histogram, then, is an overview of how many files were found in a set of mean sample value intervals. This overview can be used to trace files with large DC-offsets.

Mean distribution over all items:

Mean Occurrences
-210 - -200 : 1
-180 - -170 : 1
-160 - -150 : 4
-150 - -140 : 1
-140 - -130 : 6
-130 - -120 : 10
-120 - -110 : 2
-110 - -100 : 7
-100 - -90 : 15
-90 - -80 : 21
-80 - -70 : 21
-70 - -60 : 32
-60 - -50 : 56
-50 - -40 : 67
-40 - -30 : 202
-30 - -20 : 892
-20 - -10 : 3845
-10 - 0 : 10321
0 - 10 : 34341

```
 10 -  20 :  2703
 20 -  30 :   521
 30 -  40 :   192
 40 -  50 :   140
 50 -  60 :   214
 60 -  70 :   183
 70 -  80 :   228
 80 -  90 :    93
 90 - 100 :    65
100 - 110 :    24
110 - 120 :     5
120 - 130 :     5
130 - 140 :     8
140 - 150 :     3
150 - 160 :     3
160 - 170 :     3
170 - 180 :     6
180 - 190 :     3
190 - 200 :    10
200 - 210 :    23
210 - 220 :   173
220 - 230 :     3
230 - 240 :     6
250 - 260 :     2
260 - 270 :     1
280 - 290 :     1
290 - 300 :   104
310 - 320 :     1
370 - 380 :     1
400 - 410 :     1
440 - 450 :     1
```

Mean distribution over calls/directories:

```
Mean  Occurrences
-80 -  -70 :     2
 -70 -  -60 :     1
 -50 -  -40 :     1
 -40 -  -30 :     4
 -30 -  -20 :    14
 -20 -  -10 :    71
 -10 -    0 :   203
   0 -   10 :   671
  10 -   20 :    50
  20 -   30 :    10
  30 -   40 :     1
  40 -   50 :     1
  50 -   60 :     4
  60 -   70 :     3
```

70 - 80 : 4
80 - 90 : 3
90 - 100 : 1
100 - 110 : 2
200 - 210 : 1
210 - 220 : 3
290 - 300 : 2

None of the calls has an alarmingly high or low average sample value.

2.4 Signal to Noise Ratio

We split each signal file into contiguous windows of 10 ms and computed the Mean Square (energy) in each window. The mean sample value over the complete file was subtracted from each individual sample value before MS was computed. 5% of the windows that contained the lowest energy were assumed to contain line noise. In this way the signal to noise ratio could be calculated for each file by dividing the mean energy over all windows by the mean energy of the 5% sample mentioned above. The result was multiplied by 10*log for scaling.

SNR distribution over all items:

SNR occurrences

5 - 10 : 24
10 - 15 : 179
15 - 20 : 776
20 - 25 : 2436
25 - 30 : 4982
30 - 35 : 9094
35 - 40 : 12674
40 - 45 : 12098
45 - 50 : 7702
50 - 55 : 3403
55 - 60 : 1009
60 - 65 : 152
65 - 70 : 27
70 - 75 : 8
75 - 80 : 3
80 - 85 : 2
85 - 90 : 1
90 - 95 : 1

SNR distribution over calls/directories:

SNR occurrences
10 - 15 : 1
15 - 20 : 10
20 - 25 : 40

25 - 30 : 96
30 - 35 : 173
35 - 40 : 274
40 - 45 : 247
45 - 50 : 136
50 - 55 : 60
55 - 60 : 13
60 - 65 : 2

None of the calls has an alarmingly low average SNR-value.

5 ANNOTATION FILE

- Each line must be delimited by <CR><LF>
  OK

- Mandatory (SAM) mnemonics:
  LHD: SAM, 5.1 (or 6.0)
  DBN: SpeechDat_East_<language>_Fixed_Network
  VOL: FIXED3<LL>_<nr>
  SES: <session number>
  DIR: <with backslashes and no final backslash>
  SRC: <filename of speech file>
  CCD: <corpus code = item code>
  REP: <location of recording equipment>
  RED: <recording date, in format DD/Mmm/YYYY>
  RET: <recording time, in format HH:MM:SS>
  BEG: <begin sample, usually 0>
  END: <end sample>
  SAM: 8000 < = sampling freq.>
  SNB: 1 < = number of bytes per sample>
  SBF: < = sample byte order, meaningless with single bytes>
  SSB: 8 < = number of significant bits per sample>
  QNT: A-LAW < = quantisation>
  SCD: <speaker code>
  SEX: M/ F/ UNKNOWN
  AGE: <in years/unknown>
  ACC: <regional accent, place of growing up>
  REG: <region of call>
  ENV: <environment of call>
  NET: <network> (compulsory if mobile calls are included)
  LBD:
  LBR: <start>, <end>, [gain], [minimum value], [maximum value], <orthographic prompt>
  LBO: <start sample>, [centre sample], <end sample>, <transliteration>
  ELF: <end label file>

  OK, all mandatory labels are used.

- Optional (SAM) mnemonics (may be omitted or left empty)
  TYP: orthographic
  TXF: <name of the prompt sheet text file>
  CMT: <comment>
  NCH: 1 < = number of channels recorded>
  ARC: <region or area code of call>
  SHT: <sheet number for prompts>
  CMP: <compression, should be empty if used>
  EXP: <labelling expert>
  SYS: <labelling system>
  DAT: <date of completion of labelling>
  SPA: <SAMPA version>
  PHM: <telephone model>
  NET: PSTN < = network>
  DSC: < = discontinuity marker>
  EDU: <education level>

SOC: <Socio Economic Status>
HLT: <health>
TRD: <tiredness>
RCC: <recording conditions code>
EXT: <if >80 chars on one line>
CRP: < = corpus repetition, empty>
ASS: <assessment code>

OK, the following optional labels are used: CRP, PHM, EXP, SYS, DAT.
In addition, another set of optional labels is used, which are described in section 1.4: DIS, DIX, EPI.

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

- No illegal mnemonics used
  OK

- There are no mnemonics missing
  OK

- All files must contain the same mnemonics. This holds as well for the optional mnemonics.
  OK

- No illegal field values should appear
  OK,
  but the values of the optional files are not checked.

- Each lowest subdirectory does not refer to multiple sheet ids.
  OK

- For spontaneous speech LBR should contain a mnemonic word.
  - D1 : <date>
    OK
  - L1 : <forename_spelled>
    OK
  - O1 : <forename>
    OK
  - O2 : <city>
    OK
  - Q1 : <yes_question> or <no_question>
    OK
  - Q2 : <yes_question> or <no_question>
    OK
  - T1 : <time>
    OK

- Assessment of speech items in terms of SNR, presence of additional noise, adherence to prompting text is provided (optional)
  Not provided

6 LEXICON

- Check lexicon existence (\TABLE\LEXICON.TBL)
  OK

- The entries should be alphabetically ordered
  OK

- Used SAMPA symbols are provided in \DOC\SAMPALEX.PS
  OK

- In transcriptions only SAMPA symbols are allowed
  OK, provided that /x/ is regular SAMPA

- All SAMPA phoneme symbols should be covered
  OK

- Phoneme symbols must be separated by blanks
  OK

- A line in the lexicon should have the following format
  <grapheme form> <TAB> [<frequency> <TAB>] <phoneme transcription> [<altern.>]
  [TAB] is ASCII 9.
  OK

- Each line is delimited by <CR><LF>
  OK

- All entries should have at least one phone transcription
  OK

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

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

- Words with * or ~ should not appear in the lexicon
  OK

- The lexicon should be complete
  - Check for undercompleteness (are all words in lexicon)
    OK
  - Check for overcompleteness
    (Undercompleteness is worse than overcompleteness. Overcompleteness cannot be a reason for rejection)

We found 51 words in the lexicon which are not in the orthographic transcriptions of the database.

- Lexicon contents should be taken from actual utterances (from LBO), so the entries should exactly match the transcriptions.
  OK

- Optional information: stress, word/morphological/syllabic boundaries.
  But, if provided, then it should follow the SpeechDat conventions.
  Not provided.

7 SPEAKERS

– Check existence speaker database file (SPEAKER.TBL) or, alternatively, the session table file (SESSION.TBL). For SESSION.TBL see next section.
OK, both SESSION.TBL and SPEAKER.TBL are provided.

– Obligatory information in SPEAKER.TBL:
- unique number (speaker/caller) SCD (SPEAKER.TBL only)
  OK
- sex SEX
  OK
- age AGE
  OK
- accent ACC
  OK

– Optional information:
- height HET
- weight WET
- native language NLN
- ethnic group ETH
- education level EDL
- smoking habits SMK
- pathologies PTH
- socio-economic status SOC
- health HLT
- tiredness TRD

Not used

– Each line is delimited by <CR><LF>
OK

– Each field is separated by [TAB] (ASCII 9)
OK

– The properties of each speaker should be unique
OK, in the sense that every speaker has a unique SCD.

– Balance of sexes
- How many males, how many females, should match specification in documentation file
- Misbalance may not exceed 5% (Each sex must be represented between 45-55%)

OK
The following speaker sex distribution was found by automatic inspection of the label files:

F: 526

M:526

This is in line with the information in SPEAKER.TBL, SESSION.TBL and section 5.2 of DESIGN.DOC.

- Balance of dialect regions
  - which dialect regions and how many of each should match specification in documentation file
  - ACC or REG is used to check dialect balance, according to motivation in DESIGN.DOC
  - each region should be represented in the distribution of the country population with a deviation of 5% at the maximum and a minimum representation of 5% of the calls foreach region

OK, ACC is used for the registration of accent.
The following speaker accent distribution was found by automatic inspection of the label files:

CM: 222
CWNB: 440
EM: 79
S: 72
SB: 227
UNKNOWN: 12

This is in line with the information in SPEAKER.TBL, SESSION.TBL and section 5.1 of DESIGN.DOC. Comparison with section 5.1 also shows that the accent distribution deviates less than 5% from the population targets.

- Balance of ages
  - which age groups and how many of each should match specification in documentation file
  - Criteria
    < 16 : >= 1% strongly recommended
    16-30 : >= 20% mandatory
    31-45 : >= 20% mandatory
    46-60 : >= 15% mandatory
    (The age criteria are meant for the whole database; they are not to be applied for male and female speakers separately)

OK
The following speaker accent distribution was found by automatic inspection of the label files:
00-15 : 20
16-30 : 490
31-45 : 238
46-60 : 230
61-99 : 71
other : 3

8 RECORDING CONDITIONS

- Check existence (optional) recording conditions table (\TABLE\REC_COND.TBL) or session table (\TABLE\SESSION.TBL)
  OK, SESSION.TBL is supplied.

- Obligatory attributes of the REC_COND.TBL file (if provided):
  - recording conditions code RCC
  - region of call REG
  - environment ENV

  Not provided

- Obligatory attributes of the SESSION.TBL (if provided)
  - Session code SES
  - Recording date RED
  - Recording time (of first item) RET
  - Speaker code SCD
  - Speaker age AGE
  - Speaker sex SEX
  - Speaker dialect region ACC
  - Calling region REG
  - Calling environment ENV
  - Telephone network (if mobile calls included) NET

  OK, also PHM is included in the table.

- At least 2% of the calls must be from a public place (check ENV)

  OK
  The following environment distribution was found by automatic inspection of the label files:

  BOOTH: 21
  FACTORY: 11
  HOME: 819
  OFFICE: 137
  OTHER: 3
  PAYPHONE: 16
  PUBLIC: 7
  STREET: 16
  UNKNOWN: 22

  This is in line with the information in SESSION.TBL and section 7.1 of DESIGN.DOC.
  We consider the following environments as public: BOOTH, STREET,FACTORY, PUBLIC.

- A maximum of 5% of the calls may be made over the mobile network (the NET attribute must be used to identify them)

  OK
  The following network distribution was found by automatic inspection of the label files:

  FIXED: 1045
  MOBILE: 7

  This is in line with the information in SESSION.TBL and section 7.2 of DESIGN.DOC

9 TRANSCRIPTION

A. Validation by software tools

- Transliterations is case-sensitive unless specified otherwise.
  (In general lower case is used also at sentence beginning Only exception: proper names and spelled words, ZIP codes, acronyms and abbreviations.
  In the latter case blanks should be used in between the letters.)
  OK, transliterations are in small letters, except for the spelled letters.

- Punctuation marks should not be used in the transliterations
  OK

- Digits must appear in full orthographic form
  OK

- In principle only the following symbols are allowed to indicate non-speech acoustic events:
  [fil] [spk] [sta] [int]
  Other symbols (and language equivalents) must be mentioned in the documentation

  OK

- Asterisks should be used to indicate mispronunciations
  OK

- Double asterisks should be used for not understandable parts
  OK

- Tildes should be used to indicate truncations
  OK

B. Validation by a native speaker of the language

This validation was carried out by taking 1000 short items and 1000 long items.
The transcriptions in the label files for these samples were checked by listening to the corresponding speech files and correcting the transcription if necessary. In case of doubt nothing was corrected.

This check was performed by a native speaker of the language.

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 evaluation comprised the following guidelines:
  - Two types of errors were distinguished: speech and non-speech transcription errors
  - Non-speech refers to [fil] [spk] [sta] [int] only
  - For non-speech all symbols were mapped to one during validation.
    i.e. If a non-speech symbol was at the proper location then it was validated as correct (regardless if it was the correct non-speech symbol or not). The only exception is [sta] which should be properly marked in the transcriptions.
  - Only noise deletions in the transcription were counted as wrong, not noise insertions.
  - The given transcription is given the benefit of the doubt; only obvious errors are corrected.
  - Errors were only determined on item level, not on word level
  - 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.

RESULTS

1. Long items
Transcription errors with respect to speech were found in 34 items. This amounts to 3.4%, which is below the criterion of 5%.

Errors in the transcription of non-speech were found in 9 items. This amounts to 0.9% of the items, which is well below the criterion of 20%.

2. Short items
Errors with respect to the transcription of speech were found in 21 items. This amounts to 2.1%, which is below the criterion of 5%.

Errors in the transcription of non-speech were found in 11 items. This amounts to 1.1% of the items, which is well below the criterion of 20%.

3. Overall result

When we take the long and short item sets together then we find errors with respect to the transcription of speech in 55 items. This amounts to 2.8%, which is below the 5% criterion.

Errors in the transcription of non-speech were found in 20 items. This amounts to 1.0% which is well below the 20% criterion.

4.Further remarks

Pronunciation errors were relatively often incorrectly transcribed, e.g.:

A30131A4:
Transcribed text:
** vymazat [int]
Corrected text:
*vymazat vymazat [int]

10 SUMMARY

The Czech database was validated and approved by the SpeechDat consortium.

As serious deviations we noted:

- W1-4: We found 34 words occurring more than 5 times in the database.

This deviation was accepted by the SpeechDat-East consortium, since the criterion for the minimum number of tokens per phoneme for the phonetically rich words, is also met if the additional words would be removed from the database. Therefore, the deviation only implies additional speech data.

Below we give a brief overview of our findings for this database. The subsections follow the order of the various topics in the previous sections of the report.

1. Documentation

The DESIGN.DOC file contains all relevant information.

2. Database structure and file names

The database is formatted according to the specifications of SpeechDat-East.
Some small deviations are reported in section 2.

3. Items

The database contains all mandatory corpus items and in sufficient quantities.
The following deviation from the contents of some of the items was observed:

- W1-4: We found 34 words occurring more than 5 times in the database.

This deviation was accepted by the SpeechDat-East consortium (sse above).

4. Sampled data files

The speech data files are in the correct format (A-law). A file with acoustic characteristics of each file is delivered (SAMPSTAT.TXT). Histograms of a number of acoustic characteristics of the files (duration, mean sample value, clipping rate, SNR) were generated and included in section 4 of this report. Acoustical details of individual files can be looked up in the SAMPSTAT.TXT file.

Six sessions had a high average clipping rate. We recommend to use them only as test directories.

5. Label files

All obligatory labels are used and have the correct contents.

6. Lexicon

The lexicon is fine.

7. Speakers

The database has correct speaker balances for sex, age and accent.
SPEAKER.TBL has the correct format and contents.

8. Recording conditions

The recording conditions are in agreement with the SpeechDat-East specifications.
The SESSION.TBL file has the correct format and contents.

9. Transcription

When we take the long and short item sets together then we find errors with respect to the transcription of speech in 55 items. This amounts to 2.8%, which is below the 5% criterion. Errors in the transcription of non-speech were found in 20 items. This amounts to 1.0%, which is well below the 20% criterion.