

SPEX QQC REPORT

TITLE DATABASE: TED Phone (Translanguage English Database)
 DATABASE OWNER / PRODUCER: University of Munich, LIMSI-CNRS
 ELRA CATALOGUE NUMBER: S0032

AUTHORS OF QQC REPORT: Henk van den Heuvel
 DATE: 28 June 2002
 VERSION: 1.0

SUMMARY SHEET:

Database part	Applicability (y/n)	Quality value		
		*	**	***
1. Documentation	Y	*		
2. Format	Y		**	
3. Design & contents	Y	*		
4. Speech signals	Y		**	
5. Annotation files	Y	*		
6. Speakers	Y	*		
7. Environments	N			
8. Transcriptions	N			
9. Lexicon	N			

For each applicable part a star assessment is given.

1. * This value is given if there is not a proper and reliable fit of the contents of the SLR and the information about this part as presented in the documentation.
2. ** This value is given if the documentation well accounts for the contents of the SLR. Some small deviations are permitted.
3. *** This value is given if there is no mismatch between the documentation and the contents of the SLR.

1 Quick Quality Check Report

1.1 Documentation

The most important topics should be covered and clearly described in the documentation:

The documentation is in the following files in the DOC directory: FILE.DOC, PROMPLIST.DOC, SHORTEN.DOC, TED.DOC.

- db layout and media

OK, see files: README, FILE.DOC

- application potential for the SLR

This is not clear.

- directory structure and file names

README, FILE.DOC explain file names.

The directory structure is unclear. The first level below data are subdirectories consisting of two characters. There is no explanation as to their meaning. (It appears to be identical to the speaker code).

- recording equipment

Not documented.

- design and contents of the recordings

Contents of the recordings are in PROMPLIST.DOC. The design is unclear (how many of each type of prompts required and recorded).

- coding and format of the speech files

It is unclear what the coding of the speech files is. Unexpectedly for telephone speech, it is not A-law, but linear (16 bits, 8000 Hz). The full sessions, however, have a sampling frequency of 16 kHz AND a swapped byte order! These full sessions contain a lot of energy above 3500 Hz, which is awkward for telephone recordings.

- contents and format of the label files

Not documented.

- speakers

Not documented. There are 64 speakers. For 25 speakers the whole sessions are included both as 16 kHz speech and as laryngograph signal. The speakers should be a subset of the TED database. But the speaker initials were recoded so that they cannot be traced in the TED database. There is no text file with information about the speakers.

- recording environments distinguished

N/A

- transcription conventions

There are no transcriptions!

- lexicon: format and transcriptions included

N/A. There is no lexicon.

1.2 Format

- The file names and directory structure should correspond to the documentation

The documentation is unclear.

1.3 Design and contents

- All mandatory items according to the documentation should be included

This is unclear. There are 64 speakers. There is no list of missing files.

- Number of effectively missing files per corpus item should be appropriate

The number of recorded utterances/files per speaker varies largely, between 22 and 54.

In directory DATA/MG we find the full session (both speech and laryngograph) for speaker MG, and further recordings of speaker MF. Something went wrong here.

For 25 speakers the whole sessions are included both as 16 kHz speech and as laryngograph signal. It is unclear what this has to do with telephone recordings.

1.4 Speech signals

- For 2 CDs of the SLR acoustic measurements on the speech files will be made, and the results reported. The acoustical measurements involved are:
 - o Clipping rate
 - o SNR
 - o Mean amplitude

Acoustic values were computed for the phone recordings only (not for the full sessions nor for the laryngograph recordings). Averages were computed per speaker.

The clipping rates (based on extreme (min./max.) sample values in the files) were below 0.2% for all speakers.

The SNR-histogram over the files looks as follows:

SNR range	Number of speakers
20 - 25 :	1
25 - 30 :	25
30 - 35 :	37
35 - 40 :	1

The average mean sample values were between -300 and 300 for all speakers.

Thus, none of the speakers has alarming average acoustic characteristics.

1.5 Annotation files

- A random selection of the annotation/label files will be checked. They should be
 - o Readable
 - o Contain the information described in the documentation

The information in the SAM files is not described in the documentation. An example of a SAM label file is:

```
LHD: V1.0
FIL: speech
DBN: TED
DIR: .
SRC: c9itsnas.pes
```

SAM: 8000
BEG: 0
END: 79616
RED: 21/9/93
REP: Berlin_Eurospeech93
SNB: 2
SBF: 10
SSB: 16
RCC: 1
NCH: 1
LBD:
ELF:

All SAM files contain the same SAM-labels. The values of SAM and SBF are different for the full session recordings (16000 Hz and 01, resp.). There is no label to distinguish mother tongue and English utterances. All sessions are claimed to be recorded on 21 Sep. 1993.

1.6 Speakers

- Speaker distributions should be in agreement with documentation

Speaker information is missing.

1.7 Environments

- Environment distributions should be in agreement with documentation

N/A

1.8 Transcription

- how many speech files miss an orthographic transcription?

There are no orthographic transcriptions. For read speech the prompted texts can be derived from the file names. This is not possible for the spontaneous items.

- All non-speech markers should be described in documentation

N/A

1.9 Lexicon

- The correct set of phone symbols should be used (according to documentation)

N/A. A lexicon is not provided.

- All words in the (orth.) transcriptions should be present in the lexicon

N/A

1.10 Other remarks

None

2 Recommendations

- Transcripts of all speech files should be created. Otherwise the data cannot really be used.
- A lexicon with phonemic transcriptions of each word would be desirable.
- A better documentation with information about the speakers and a link to the speakers in TED is highly desirable.