Specification of validation criteria

Validation criteria for Nemlar Arabic TTS database

Niklas Paulsson (ELDA)

2005

The NEMLAR project is supported by the INCO-MED programme

http://www.nemlar.org

Contents

1 Executive summary

This document contains the specifications of the validation criteria for the Arabic TTS database within the Nemlar project. This document gives an overview of the aspects of the database which are validated, such as, e.g. documentation, completeness of database items or acoustic quality of the speech signal. It presents a set of tolerance margins, otherwise called validation criteria, which are employed to accept or reject a database. It also outlines the validation procedure which is done in a number of stages.

2 Introduction

The aim of this document is to specify validation criteria that Nemlar databases should fulfill and give an overview of the aspects of the databases that are checked in the process of validation. This document lists the criteria against which the databases are checked and which are employed to accept or reject a database.

The principles of validation and the criteria listed here have evolved over a number of previous TTS projects. Arabic languages we are dealing with in Nemlar, pose new challenges that have to be accounted for during validation as well.

Apart from very specific validation criteria (like the allowed number of missing files) the databases should also fulfill 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 fulfillment is of immediate importance for the acceptability of a database.

The following aspects of the validation criteria are addressed in the sections below:

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. Database design and collection.
8. Recording conditions.
9. Transcription quality.

3 Documentation

Each database must be accompanied by a DESIGN.DOC which is written in English and 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 speech and 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
  - baseline corpus & components
  - voice conversion corpus & components
  - expressive speech corpus & components
- recording platform description
  - microphone positions
  - microphone types
  - laryngograph
- speaker demographics information:
  - speaker selection
  - accent regions of each speaker
  - a reasoned description of the regional pronunciation variants that are distinguished
  - age groups of each speaker
  - sexes: males, females (also children) of each speaker
- orthographic transcription 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.)
  - annotations 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
- prosodic transcription information:

  - procedure for prosodic transcriptions
  - conventions for prosodic transcriptions
- phonetic transcriptions information
  - procedure for phonetic transcriptions
  - conventions for phonetic transcriptions
  - conventions for phonetic segmentation
- lexicon information:
  - procedures to obtain phonemic forms from orthographic input
  - list of SAMPA phone symbols (http://www.phon.ucl.ac.uk/home/sampa/arabic.htm)
  - list of PinYin 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 subcorpora.
  - 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.

4 Database Structure, File Names and Contents

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

The databases should comply with the following directory structure:

```
/<database>/<Type>/<subcode>
```

Where:

<database>	Defined as <dbName><#><language code> where <dbName> is NMTTS, <#> is 7 for NEMLAR, <language code> is <ISO-639 code>: FOR OUR DATABASE: NMTTS7AR The code follows the convention to have the language code part (ISO 639-2 standard) in lowercase, the country code part (ISO 3166 standard) in uppercase and the third 3-8 letters subtag in lowercase, e.g. SINCE THE SPEAKERS WILL BE EGYPTIANS: AR, ara, EG for arabic and egypt.
<type>	Defined as <TTT>

	Where <TTT> is the type of the subcorpus, i.e. either BLR (Baseline),
<subcode>	Defined as <TTT><SS> Where <TTT> is defined as above, and <SS> is the subcorpus code, i.e. either 11 or 12 (for transcribed speech); 20 (for written text); 31,or 32, or 33 for selected phrases.

Table 1 – Directory structure

Both signal files and label files have to be put in the terminal node subdirectories.
In addition to the previous structures the following directories are used to store the other (non-speech data) files:

/	README file with overview of database, DISK.ID file, and copyright file
/<database>/DOC	Documentation
/<database>/HTML	HTML access to (selected) recordings
/<database>/TABLE	Speaker, recording condition, environment conditions, and lexicon tables
/<database>/INDEX	Index files
/<database>/PROMPT	Prompt sheet samples
/<database>/SOURCE	Source code

Table 2 – Documentation file hierarchy

The filenames should correspond to the following template:

```
<dbID><T><subcode><SS><NNNN>.<LL><F>
```

where:

<dbID>	Database Identification Code (00-ZZ), for NEMLAR TTS: T7
<T>	Type, where B is baseline,
<subcode>	Two digits corpus subcode: 11, 12, 20, 31,32, 33
<SS>	Two digits speaker ID (SCD)
<NNNN>	Utterance identification number: 0000-9999
<LL>	ISO 639 langauge Code (Here: AR)
<F>	File type code: S: SAM label file L: pitchmark file from laryngograph recordings (text) P: phonetic segmentation file (text) 1,2,3: speech signal files for channels 1 – 3

Table 3 – File name conventions

4.2 The DOC directory

The following files should be in \<database_name>\DOC:

- DESIGN.DOC
- PLATFORM.DOC
- TRANSCRP.DOC (optional)
- SPELLALT.DOC (optional)
- SAMPALEX.PS

- ISO8859<n>.PS
- SUMMARY.TXT
- SAMPSTAT.TXT
- VALREP.DOC
- TRANS-13.DTD

The validation of the DESIGN.DOC main documentation file is described in section 3. PLATFORM.DOC contains platform specifications. TRANSCRP.DOC contains 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. SUMMARY.TXT contains an overview of all items recorded for each session. SAMPSTAT.TXT is the output of the acoustical check on the speech files performed by each partner. The file VALREP.DOC which contains the validation report is created by the validation centre.

4.3 The TABLE directory

Tables should be in \<database>\TABLE

- LEXICON.TBL
- SPEAKER.TBL
- REC_COND.TBL

4.4 The INDEX directory

Index files should be in \<database>\INDEX

- CONTENTS.LST

4.5 Other directories

The root directory should contain the files:

- README.TXT: ASCII text file containing a description of the files in the database
- README.HTM: with browser access to all documentation directories (optional)
- COPYRIGH.TXT: copyright statement in ASCII
- DISK.ID: 11-character string with volume name

4.6 Other requirements

All text files should have <CR><LF> at line ends. This concerns all label files, all table (.TBL) files, all index (.LST) files, and all (.TXT) files.
All table files and index files (but *not* SUMMARY.TXT) should report the field names collected in each record as the first row (header) of the file. In this header tabs should be used to separate the fields just like in the rest of the file.
Empty files are illegal. This is of special relevance for speech and label files.
For each label file there must be one corresponding speech file and vice versa.
Obviously the 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.

Notation of Text Corpus	Kind and Size of Sub Corpora of Corpus C
C1_T consists of:	**Transcribed speech**
C1.1_T	transcribed speech
C2_T	**News, Novels and short stories with short sentences**
C3_T consists of:	**Constructed Phrases**
C3.1_T	Frequent used phrases
C3.2_T	Triphone coverage sentences

5.2 Validation of missing items

For each database it will be checked if all mandatory items are present in sufficient quantities:

- For C_T the correct minimum number of word tokens per kind of sub corpus and domain will be checked from the prompt texts. The number of word tokens can deviate by 5% per sub corpus.
- the corpus C_T should contain around 95% of the speech segments found in the common word lexicon. For each common speech segment at least 2 candidates should be present.
- 10 is the minimal number, each phoneme has to be present in C3.3_T
- 5% of the paragraphs could have less than 25 word tokens. 1% of the text C1_T can be missing in C1_PS.
- 1% of the text C2_T can be missing in C2_PS.
- 1% of the text C3_T can be missing in C3_PS.
- all the prompt sheets have to be read. 0.1% of the word tokens read may be missing.

6 Acoustic Quality of the Speech Files

During validation, the acoustic quality of the recorded speech will be checked. The speech files should met the following criteria:

- Format: 96 kHz, 24 bits, 2 channels
- SNR: at least 90% of all files should have a SNRA of more than 40 dBA.
- Frequency: 40-20000 Hz with max a deviation of 0.5 dB.
- Reverberation: has to be in the range RT60 > 0.3 seconds

Also there must be a corresponding laryngograph file for each speech file.

7 Label Files

Checks will be performed to make sure that:

- Correct labels and correct accompanying values are used
- There are no empty label files
- Each line is delimited by <CR><LF> (DOS format)

Formal correctness of SAM label files is checked. It will be checked if there is one SAM label file for each set of speech files, and if there is a set of speech files for each SAM label file.

8 Lexicon

The TTS database will contain three lexica:

- common word lexicon
- extended common word lexicon which are not included in the common word lexicon
- a small proper name 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 (i.e., mispronounced or truncated words)
- If tagging is supplied, check that all tag symbols are defined and only those symbols are used

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

8.2 Validation of phonemic transcriptions

- For the extended common lexicon the correctness of 500 phoneme transcriptions and 500 POS tags will be checked. 5% errors are allowed.
- For the small proper name lexicon the correctness of 500 phoneme transcriptions will be checked. 5% errors are allowed.

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%

Table 5: Confidence intervals

9 Recording Conditions

A platform with 2 synchronized channels is mandatory:

- channel 1: large membrane microphone
- channel 2: laryngograph

Synchronized signals recorded from a large microphone and a laryngograph has to be provided.

10 Transcription

Transcription validation of speech is carried out by a trained native speaker of the language concerned, who did not participate in the original transcription process. The transcription validation of the 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 of deciding 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 receive the benefit of the doubt and that only overt errors should be corrected.

10.1 Annotation levels

For each utterance (speech file) it is required to provide:

- the prompt text used to elicit the utterance,
- the orthographic transcription,
- the prosodic transcription,
- the segmentation into phonemes (can be performed partly automatically),
- the pitch marks, associated with the glottal closure (can be performed partly automatically).

10.2 Speaker information

The presence of the minimum amount of speakers for each voice type will be checked. During the speaker selection process the speaker's profile will be assessed.

10.3 Orthographic transcription

250 files (random sample from full db) of the orthographic transcriptions are checked. A max WER (Word Error Rate) of 0.1% is permitted. Ideally 5K words at least are validated.

Native speakers of the language perform the check on the transcriptions.

10.4 Prosodic transcription

A sample of 1000 phones will be checked for the prosodic part. A max WER of 5% of the POS tags is allowed.

10.5 Segmentation

A sample of 1000 phones will be checked for the segmentation: 500 phones from the manually segmented and 500 from the automatically segmented. A max of 5% wrong segmentations is allowed for the manual part and 10% for the automatic part. An error is defined as a deviation larger than 20 ms.

10.6 Pitch marks

The same parts as for the orthographic validation will be checked for the pitch marks. A max deviation from the reference pitch mark of 5% is allowed and that is not bigger than 0.5 ms. Also a max of 3% of unvoiced/voiced errors is allowed.

10.7 Statistical reliability

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%

Table 8: Confidence intervals for 1000 items

10.8 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.

11 Validation procedures

The Nemlar TTS database will be validated in two stages: a full validation to check the compliance with the technical criteria for the recordings themselves and in a pre-release validation to check that all documentation and files are present and correct.

12 References

[1] van den Heuvel, H.*: Validation criteria*. Orientel. Technical Report D6.2. Version 1.2, 2002.