
# Specifications of Arabic TTS Speech corpus

Khalid Choukri (ELDA)

Salah Hamid (RDI)

With the contribution
of the NEMLAR partners

3rd May 2005


The NEMLAR project is supported by the INCO-MED programme


http://www.nemlar.org

Table of Content

# I Language Resources to be produced within NEMLAR

## I.1 Type of Resources

Following the work carried out within the other work packages of NEMLAR, the consortium agreed to focus on three main resources:

(A) Written corpus
500 K words of annotated, including vowelised corpus fully packaged both the annotated and unannotated parts

(B) Speech corpus for TTS applications
Recordings/ segmentation/ etc. of 10 hours (2x5 hours) for speech synthesis , 5 hours of a Male speakers and 5 hours from a female speaker.

(C) Broadcast news
The data will consist of about 40 hours to be provided by ELDA of Arabic data (mainly Standard Arabic from a number of broadcast companies); Transcriptions will follow the Transcriber conventions as used by ELDA and focus on the orthographic, named entities, speaker/turn segmentation levels. No phonetic transcription/segmentation is planned.

## I.2 Language Resources Production within NEMLAR

Within WP5 plans the partners that are in charge of Language Resources productions are:

(A) Written corpus: this part will be carried out by RDI and Amman University (50% each). Tools will be provided by RDI. Raw data will be provided by RDI (IPR cleared and Copyright free) and ELDA.

(B) Speech corpus for TTS applications: Recordings/ segmentation/ etc. of 10 hours will be done at RDI facilities, the processing will be done by RDI for the male speaker and by ENSIAS for the female speaker.

(C) Broadcast news: Transcriptions will be provided by RDI; ELDA will support the RDI team with its expertise in Transcriber exploitation.

## I.3 Language Resources Validation within NEMLAR

Data validation will be conducted by ELDA with the support of CST and UU. Some involvement of UOB and Sotelet-IT is foreseen.

## I.4 Packaging of Language Resources to be produced within NEMLAR

Once the data is validated and a pre-release version is accepted by the validation center, each producing partner will be in charge of supplying its part to RDI who will be in charge of packaging the three sets.

The following sections are adapted from the TC-STAR Project (FP6-506738) and in particular the deliverable D8a and should remain confidential to ELDA, RDI, and ENSIAS till this document become publicly available via ECESS or TC-STAR.

# II Specifications of Language Resources for Speech Synthesis Language Independent Part … Adaptation to the requirements of the NEMLAR project on Modern Standard Arabic

The original deliverable is: **Deliverable no.: D8 and its title is: «Specifications of language resources for speech synthesis".**

**The Lead contractor for this deliverable is Harald Hoege, Siemens AG and the** Author(s) are: **Harald Höge, Antonio Bonafonte, Henk v.d. Heuvel, Asuncion Moreno, Herbert S. Tropf, David Sündermann, Ute Ziegenhain.**

# III The Rationale of the Specifications

The specification of language resources for speech synthesis has been addressed by various authors (e.g. /Ellbogen2004/, /Black2004/). According to these specifications language resources have been built for European languages. The aim of this document is to come up with specifications on language resources (LR) for speech synthesis based on which LRs in a variety of languages can be produced. The specifications are being developed within the framework of TC-STAR[1]. Within this project LRs for TTS systems and selected research areas on speech synthesis will be generated for the languages **UK-English, Spanish and Mandarin**. Furthermore, the document aims at serving as a basis for other projects like ECESS[2] which aim to cover more languages. In the context of HLT these specifications can be seen as a starting point to specify a 'basic language resource kit' (BLARK) [3] for speech synthesis.

This document aims at complementing the original one and ensuring that it is easily customised to another language: modern standard Arabic as addresses in the NEMLAR project[4].

D8a_LIP is one part of a deliverable which covers the language independent part (LIP) of the specifications of language resources for speech synthesis; it is the one that helped design the current one to which the Language specific issues and language specific deviations from the language independent specifications are described in an appendix (referred to as language specific part or LSP).

## III.1 Focus of this document in conjunction with D8a_LIP

This document describes the language independent specifications for language resources (LR) which are needed for building speech synthesis systems and for investigating specific research topics in speech synthesis. As a first attempt for Arabic and in the context of NEMLAR the LR should be suitable to **build the most advanced state-of-the-art TTS systems (at least for concatenative speech synthesis).**

[1] www.tc-star.org
[2] www.ecess.org
[3] BLARK is an initiative of the HLT community to make available needed language resources for each language
[4] See www.nemlar.org for details about the project

The creation of voices for TTS systems will be based on **read speech.** For this issue, text corpora are specified which have to be read by TWO selected speakers (one male and one female).

The main chapters of this document are:

- the construction of the text corpora to be read
- the procedure to select suited speakers
- the recording platform
- the annotation of the recordings of the speakers
- the database interchange/distribution format
-

The language resources created within the project will be **validated**. For this purpose specific validation criteria will be developed.

### III.2 Notation of Corpora

In the following paragraphs the corpora are denoted in general by Cn_xy , where

- n denotes the design principle of a certain sub corpus of the corpus C. If n is omitted the complete corpus is denoted
- x denotes the application of the corpus (e.g. corpus for voice conversion); x is not always denoted
- y denotes the content of the corpus: T (Text), PS (Prompt Sheet), R (Recorded speech) etc.; y is not always denoted

### III.3 Design Principles of the Text Corpora

The basic design principle relates to the term 'suitability'. In the context of NEMLAR the term suitability refers to LRs which are optimal for generating the most advanced state-of-the-art TTS system.

For building a general purpose TTS system, speech has to be synthesized for any application area. Application areas can be described in terms of 'domains', where the term 'domain' is defined either by a lexical field such as politics, sports and culture or by a communicative situation[5] such as 'read speech', 'conversational speech', etc. Both aspects are relevant for speech synthesis. D8a focuses on the specification of LRs derived from the communicative situation 'read speech' but also aims on designing LRs covering different domains to build TTS systems having a high coverage on all the domains relevant for the culture in a given language. A similar goal has been addressed in the EU-funded project LC-STAR[6] , where lexica with a high coverage on certain domains have been created. The domains selected in LC-STAR, serve as a basis for some of the domains described in this document.
The main issue in synthesizing speech from any domain is to achieve a good coverage on speech segments used in a given language. In the following paragraphs speech segments like triphones or syllables in various prosodic contexts are regarded. It is assumed that triphone or syllable coverage is mainly independent from domains. Although triphones or syllables are quite 'large' segments, the synthesized speech using state of the art concatenation technology still is far from 'perfect'. This drawback is due to problems in manipulation of concatenated speech segments. In order to achieve more or less 'perfect' coverage on a variety of different domains a sub corpus called 'frequent used phrases' is specified which is constructed from domains as specified in LC-STAR.

[5] Within the DARPA projects communicative situations are defined for supporting research in ASR. For this purpose LR for different communicative situations as 'read speech', 'conversational speech', 'broad cast news', 'call home' etc. are defined and related LRs are provided
[6] www.LC-STAR.com

Another issue to be accounted for is the coverage of supra-segmental prosodic events e.g. phrase breaks, phrasal and sentence accent and intonation contour, etc in the corpora to be read. This information is needed to make models for prosody and to provide speech segments with are suited to be used for all the prosodic contexts to be synthesized.
The problem of achieving high coverage on supra segmental events has not been deeply investigated. However in the specifications it is taken into account that certain linguistic structures evoke certain prosodic events. Furthermore linguistic structures found in text (e.g. as found in a newspaper) differ from those found in speech. For this purpose text derived from 'written text' and text derived from 'transcribed speech' - i.e. from speech corpora where the utterances have been converted to text - is specified.
To increase the prosodic coverage of the segments with respect to their position at the beginning and the end of a sentence, a corpus on written text containing many **short sentences** is specified additionally.

Including all the considerations made above, the NEMLAR TTS text corpora are composed by a set of the following sort of sub corpora is regarded:

- C1: transcribed speech (if this is possible to re-use some of the transcribed speech from different domains, being produced in the Broadcast news Task)
- C2: news excerpts, novels and short stories with short sentences (written text from different domains)
- C3: constructed phrases (text specifically constructed)

C3'*constructed phrases*' is composed by 3 sub corpora:

- C3.1: 'frequent phrases': serves to improve the quality of frequently used phrases like phrases build from dates, numbers, yes/no expressions and for frequent used phrases found in domains as defined in the LC-STAR specifications.
- C3.2: 'triphone coverage sentences': serves to improve the coverage with respect to missingtriphones or syllables.

### III.4 Size of the Text Corpora

For building voices for TTS systems from corpora the recorded corpora should have a good coverage on the basic speech segments together with their prosodic properties. It is evident that the higher the amount of recorded speech the better should become the coverage. However a compromise between coverage and effort in creating the LRs has to be made.

**For building a single voice in a given language for a state of the art speech synthesis system a total volume of** 5h **of speech is considered to be adequate.** This assumption is endorsed by the NEMLAR project.

Assuming 0.4 sec duration in average per word 10 h of speech corresponds to the time needed to read a text corpus of about 90 000 words. This amount of words is distributed on the sub corpora described above according to some balanced statistics:

- **C1_T transcribed speech** ***6600 per speaker***

- **C2_T: written text** ***16500 per speaker***
- **C3_T: constructed phrases** ***9900 per speaker***

**consists of:**

  - C3.1_T: Frequent Phrases ***3300 per speaker***
  - C3.2_T: Triphone Coverage sentences ***6600 per speaker***

The complete corpus of the text corpora C1_T, C2_T, C3_T is denoted by C_T and is also called the '***Baseline Text Corpus***'.

**In order to get a good coverage on the speech segments covering a spoken language the reference corpora of C1 and C2_T have to be derived from much larger corpora.**

### III.5 Building Voices and Related Recorded Corpora

Based on the text corpus C_T prompt sheets are made resulting in the corpus C_PS. Using the prompt sheets different speakers are recorded to produce different 'voices'.

#### *III.5.1 Voices and Related Corpora for Building TTS-Systems*

The voices recorded to build a TTS system (baseline system) are called '***baseline voices***'. In general, for the baseline system 1 male and 1 female voice will be recorded and this is what was agreed within NEMLAR. Given the amount of text as defined in III.4, the duration measured in hours (h) will be for each voice and for each language approximately:

- C1_BLR: recorded transcribed speech ***1h***
- C2_BLR: recorded written text ***2.5h***
- C3_BLR: recorded constructed phrases ***1.5h***

The resulting corpus C1_BLR, C2_BLR, C3_BLR is called the '***Baseline Recorded Corpus***' (**C_BLR**), which contains about 10h of recorded speech.

#### *III.5.2 Speaking Mode*

Speakers should read the text in a manner that good results concerning the quality of the speech synthesis system as well as the suitability for research are achieved. Ideally the recordings should cover different speaking modes and the speech segments should cover all phonetic variations as well as all prosodic variations and all kinds of speaking modes. According to current state of art of concatenative speech synthesis systems, the concatenation of speech segments selected from corpora with different speaking style and expressivity leads to unsatisfactory results[7]. Due to the restriction of the corpus to 5h speech recorded from one speaker it was decided to focus mainly on coverage of phonetic and prosodic variations with speakers speaking in a rather neutral manner.

#### *III.5.3 Selection of the Speakers and Related Corpora*

The selection of the base line speakers is done very carefully. Selection criteria are pleasantness of the voice and the suitability for speech synthesis based on concatenation and pitch synchronous manipulation. A specific procedure for selection is defined.

### III.6 Studio for Recording, Speech Quality and Pitch Marking

The usefulness of the recorded speech depends on the quality of the speech signal and on the precision with which the glottal closure can be reliable marked.

The quality of the speech signal is defined by the parameters:

[7] If the speech segments concatenated are from speech sections with different speaking modes, the synthetic speech sounds not 'consistent': Remark from Nick Campbell on the second ECESS meeting (Maribor June2004)

- signal to noise ratio of the recorded speech
- bandwidth of the speech signal.

For supporting the marking of the glottal closure with the requested precision a laryngograph has been proven useful though not all persons deliver a useful signal from the laryngograph. This fact has to be taken into account when selecting the speakers.
To precisely locate the position of glottal closure the reverberation of the room should be as low as possible.
For making research in the radiation of speech optionally stereo recordings could be made. This approach supports the fact, that the wave sources radiated are partly uncorrelated for different directions.

### III.7 Annotation

Annotation and segmentation is based on the following rules:

- All speech recordings are transliterated in normalized text form using Arabic vowelized text scripts.
- All speech transcriptions are tagged (POS) and annotated with specific markers, which are important for selecting speech units (e.g. noise, unintelligible words, etc)
- All speech recordings have to be marked prosodically.
- For baseline voices the speech recordings are completely phonetically transcribed and manually checked listening to the real recordings.
- For baseline voices the speech recordings are completely segmented in phones/syllables on signal level. 2 h of speech are checked manually.
- For baseline voices the speech signal of the speech recordings is completely pitch marked. 2h of speech are checked manually.

### III.8 Database interchange format

The core concept used in SpeechDat/SpeeCon databases (i.e. that a metafile should contain, in a promptly accessible form, redundant information that is also present in the label files for individual recordings) is maintained. Two types of label files are used: 1. a SAM file containing general information pertaining to the corresponding speech file as a whole (including the complete transcriptions), the complete recording of the speaker and the complete database; 2. label files for each corresponding speech file containing time stamped information (pitch marks, phonetic segmentations). A similar approach was also followed in the NETWORK-DC project (http://www.hltcentral.org/projects/detail.php?acronym=NETWORK-DC).
LC-STAR standards were followed for the TTS-lexicon and for the categorisation of domains in C3.3.

### III.9 Validation Criteria

The validation criteria are integrated in the document. They are marked by:
*'Validation: description of the validation criteria'; they have to be agreed upon between RDI, ENSIAS and ELDA.*

### III.10 Languages

Within NEMLAR only Modern STANDARD Arabic is to be covered. Applications concerning colloquial Arabic may be conducted by the partners outside the project using the same database.

For other projects as ECESS (http://www.ecess.org/), other languages will be added.

### III.11 Number of Speakers

The number of speakers for generating voices is specified per language.

| Number of speakers | Kind of voice |
|---|---|
| 1 | Baseline voice male |
| 1 | Baseline voice female |
| Status | Mandatory if not specified otherwise in D8a_LSP |
| Recommendation | |
| Comment | |

Validation: the presence of the minimum amount of speakers for each voice type will be checked.

### III.12 Speaker Profile

| Feature | Native / bilingual speakers |
|---|---|
| Native skill | **For baseline voices:**<br>The speaker for a given language has to be a native speaker of that language, and, without any doubt, the given language has to be the speaker's dominant language. |
| Status | Mandatory |
| Recommendation | For baseline voices: Both parents of a speaker should be fluent speakers of the given language as well. |
| Comment | |

| Feature | Age of speakers |
|---|---|
| Value | 22 – 50 |
| Status | mandatory |
| Recommendation | |
| Comment | |

| Feature | Speaker experience |
|---|---|
| Value | The speaker has to be a professional speaker, e.g. newscaster, announcer, narrator, reciter, actor/actress with elocution classes, etc. |
| Status | mandatory |
| Recommendation | Elocution classes and being active in that profession |
| Comment | |

*Validation: During the speaker selection process the speaker's profile will be assessed.*

### III.13 Speaking Modes

Speaking style

| Feature | Speaking style |
|---|---|
| Value | fluent reading compatible with a newscaster/announcer speaking in neutral manner |
| Status | Mandatory |

| | |
|---|---|
| Recommendation | |
| Comment | no (imitated) spontaneous speech, no dialogue style |

Voice Quality

| Feature | Voice quality |
|---|---|
| Value | Pleasant;<br>consistent, even, uniform voice quality produced by each speaker throughout all sessions |
| Status | Mandatory |

Expressivity

| Feature | Expressivity |
|---|---|
| Value | The dominant expressivity is that chosen by the speaker compatible with a professional newscaster/announcer speaking in neutral manner. |
| Status | Mandatory |
| Recommendation | |
| Comment | |

### III.14 Casting of speakers

Selection of Speakers for the Baseline Voices

***Validation after the pre-selection phase****: the pre-selected speakers will be assessed by auditory inspection on native tongue, age and proficiency (by trained native speakers of the language). A short speech sample of each selected speaker is submitted to the validation centre. Here the validation centre has only an advisory role.*
***Validation after the final selection phase****: the voice quality based on synthesized sentences of the final selected speakers will be assessed. The 10 synthesized sentences of each selected speaker are submitted to the validation centre. Here the validation centre has only an advisory role.*

## IV Detailed Specification of Corpora

Within this adapted LIP document in total THREE (3) corpora are specified:

- corpus for the pre-selection of the baseline voices
- corpus for the final selection of the baseline voices
- corpus for the creation of baseline voices

All corpora consist of recorded read speech, where the speakers read all or selected parts of a text corpus C_T as defined in section IV.2

### IV.1 Notation of Sub Corpora C

The general notation of a corpus is C.
In the following the corpora are denoted by Cn_xy, where

- n denotes the domain of a certain sub corpora of the corpus C; if n is omitted the complete corpus is denoted
- x denotes the use of the corpus (e.g. corpus for voice conversion); x is not always denoted

- y denotes the content of the corpus: T (Text), PS (Prompt Sheet), R ( recorded speech) etc.; y is not always denoted

## IV.2 Kind and Size of Sub Corpora of Corpus C_T

The text corpus C_T is composed out of the sub corpora as documented in the table below.
The corpora C1.2_T, C2_T, and C3.2_T have been designed to achieve high coverage with respect to speech segments and prosody.
The text corpus C1.1_T is a text corpus of transcribed speeches from the NEMLAR broadcast news task. The corpus C3.1_T has been designed to cover frequent used expressions. The corpus C3.3_T is defined to contain sentences with high phoneme/syllable coverage including rare phonemes/syllables.

| Notation of Text Corpus | Kind and Size of Sub Corpora of Corpus C |
|---|---|
| **C1_T consists of:** | **Transcribed speech** [8] |
| 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 |
| | |
| Status | mandatory |
| Recommendation | |
| Comment | |

### *IV.2.1 Domains and Size of Corpus C3.1_T 'Frequent used Phrases'*

The domains D1 – D5 chosen were used in LC-STAR to define the domains for the common word lexicon. Frequent used phrases have to be designed as derived from written text (article, news paper, etc.). D0 is a domain which covers the 'very most' frequent expressions. The related phrases should be designed as coming from transcribed colloquial speech.
The phrases are embedded into sentences. Within a single sentence expressions from the same or several sub-domains could be integrated with the goal to achieve high coverage for all domains. The amount of text (measured in numbers of words of a text) dedicated for the different domains should serve as an indication. As the kind and number of very frequent used phrases depends on the culture in which a given language is spoken the concrete figures of the amount of words per domain is documented in the language specific documentation D8a_LSP based on the following recommendations.

For NEMLAR we may inspire from this and from the data used within the Orientel project that can be provided by ELDA. In particular D2.1.2 and the other specification documents referred to as:

| | |
|---|---|
| D2.1 | Speech database design |
| D2.1.1 | Specification of recording scenarios and platform requirements |
| D2.1.2 | Specification of corpus and vocabulary |

[8] From the Nemlar broadcast news task.

| **Domains** | **Sub-Domains** | **Further descriptions In English , localized version should be considered, in addition to other cultural aspects:** | **Size in words TO BE ADAPTED** |
|---|---|---|---|
| D0. frequently used colloquial expressions | D0.1 numbers (cardinal and ordinal)[9] | 0,.., 10, tenth, hundred, ..., billion and all numbers which cannot be derived by concatenation rules (e.g. eleven, twelve, first, second,...) | 1750 |
| | D0.2 measures of length, weight, time, content, temperature | Inch, pound, hours, barrel, Fahrenheit | |
| | D0.3 all dates from 1 – 31, all days of the week dates, all months, time expressions, years around 2005, special days | 2004-12-29; 6pm; 1996; Christmas; Monday; March ***also Hijiry calender month names and Arabic names of gregorian calender month names should be included.*** | |
| | D0.4 seasons; time expressions | **Spring; the day after tomorrow** | |
| | D0.5 abbreviations | **.com, .info, .net, .info, .org; EU; EC; TC-STAR; IBM; appreviations are not used *in Arabic language abbreviation usage is not common and if forign language abbreviations are contained in arabic text it is pronounced as transletrated words (e.g. "ای بی اِم")*** | |
| | D0.6 world wide important cities and countries | Paris, Beijing, Germany | |
| | D0.7 signs | #, !, $, | |
| | D0.8 home; Kitchen | Bathroom; coffee machine | |
| | D0.8 materials | Iron; glass; paper | |
| | D0.9 operate tools | Stop; switch on; | |
| | D0.10 yes/no expressions | OK, Not at all; | |
| | D0.11 Common parts of body | Leg, brain | |
| | D0.12 Common illness | influenza | |
| | D0.13 Greetings | Good morning, hello | |
| | D0.14 some international known names | Siemens, Beethoven, Schröder | |

[9] all 'root expressions' building numbers should be included together with their prosodic variations ( e.g. prosody with respect to the position (end, middle, beginning) of a composed number)

| | | | |
|---|---|---|---|
| | D0.6 miscellaneous | Sorry, you are welcomed, thanks, | |
| D1. Sports/Games | D1.1.Sports (special events) | soccer, skating, skiing, tennis, baseball, betting, lottery etc. | 350 |
| D2. News | D2.1. Local and international affairs | top stories on domestic and foreign affairs, headlines with articles, etc. | 800 |
| | D2.2. Editorials and opinions | special reports, article of the day | |
| D3. Finance | D3.1. Business, domestic and foreign market | articles on stocks, currencies, earnings, articles on transactions, articles on companies etc. | 350 |
| D4. Culture/Entertainment | D4.1. Music, theatre, exhibitions, review articles on literature | articles/reviews (no primary literature) on musicals, shows, comedies, movies, theatre, art, TV-shows, etc. | 350 |
| | D4.2. Travel / tourism | description of regions/ surroundings, sites, more general descriptions of specialties of local cuisine, etc. | |
| D5. Consumer Information | D5.1. Health | articles on health for non-specialists | 350 |
| | D5.2. Popular science | articles for lay people | |
| | D5.3. Consumer technology | Descriptions & manuals of mobile phones, PDA's, TV, video recorder, etc. | |

*Validation: 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.*

### *IV.2.2 Coverage Issues of the Text Corpus C_T*

Part of the corpus C_T has to be designed in order to achieve high coverage with respect to the speech segments of a language. This issue concerns the text corpora C1 _T, C2_T, and C3.2_T. The other text corpora C1.1_T, C3.1_T and C3.3_T are designed according to other criteria.

#### IV.2.2.1 Definition of Speech Segments

In this document the speech segments of a given language are either triphones or syllables[10]. The set of speech segments used is documented in D8a_LSP. Each segment is defined by its symbolic annotation and by its prosodic properties. Following prosodic properties are differentiated:

**Prosodic property on position:**

A speech segment can have different positions within a phrase:

- Initial of a prosodic phrase
- Middle of a prosodic phrase

[10] should be adapted to Arabic (diphone for spanish, syllables for Mandarin)

- End of a prosodic phrase,

For the end position 3 different cases have to be distinguished:

- End-statement
- End-question
- End-exclamation
- End- questinal exclamation ("?!" after phrase mark)
- End-'more-coming' (typically end of a phrase).

**Prosodic property on stress for languages using stress:**

As denoted by a canonical TTS lexicon a speech segment can have the stress modes

- stressed
- unstressed.[11]

A more detailed definition of the prosodic properties can be found in the appendix.

***Stress in Arabic language has two types.***

- ***Word stress(each word has one stressed syllable)***
- ***Phrase stress in which all syllables within a selected word of the phrase will be stressed (it depends on meaning )***

#### IV.2.2.2 Specification of Segment Coverage

Symbolic Coverage

The speech segments found in the baseline text corpus should cover a high percentage of the speech segments found in the LC-STAR common word lexicon if available. If not we should aim at the best coverage possible with the phonetic set of modern standard Arabic. The value achieved for speech segment coverage is documented in the LSP.

Validation: the corpus C_T should contain around 95% of the speech segments found in the TC-STAR common word lexicon. For each common speech segment at least 2 candidates should be present.

#### IV.2.2.3 Achievement of Segment Coverage

The method to construct the corpora is not mandatory. It is a 'best practice' proposal.
Within the corpus C the sub corpora C1.2_T and C2_T can be constructed starting from larger corpora (the other sub corpora are constructed according other principles). To keep the effort within limits, the starting corpus should not be larger than 20 - 100 times than the target text leading to the table below:

| Notation of Sub Corpora | Sort of text and quantity in number of words for **starting texts** to derive the sub corpora C1.2_T and C2_T |
|---|---|
| C1.2_T | general transcribed text in principle 1M word tokens |
| C2_T | novels and short stories with short sentences 1M |
| Status | Mandatory |
| Recommendation | |
| Comment | |

[11] stress does not exits in languages like Mandarin; this issue will be adressed in D8a_LSP to adapt this feature to Arabic.

#### IV.2.2.4 Phoneme Coverage

The used set of phonemes including rare phonemes is documented in D8a_LSP.
The corpus C3.3_T has to yield a high coverage on phonemes including rare phonemes. The algorithm to achieve high coverage is described in Appendix A.

*Validation: 10 is the minimal number, each phoneme has to be present in C3.3_T*

#### IV.2.2.5 Prosodic coverage

The prosodic aspects will be handled in a way to ensure that all speech segments are found in all positions.

Validation: the values X and Y are inspected 5% deviations from the values as denoted in the LSP are accepted.

### *IV.2.3 Prompt Sheets Corpus C_PS*

The prompt sheet corpora C1_PS , C1.2_PS, C2_PS, C3.1_PS, C3.2_PS and C3.3_PS are derived from the text corpora C1_T, C1.2_T, C2_T, C3.1_T, C3.2_T and C3.3_T.
The corpus of all prompt sheets is called C_PS

#### IV.2.3.1 C1_PS (Transcribed speech)

The prompt sheets are derived from C1_T. The complete text corpus C1_T must be used for C1_PS. The text is grouped into small paragraphs. The paragraphs should be as small as possible. But they should be large enough to achieve a most natural prosody as observed if a complete text would have been read. Each paragraph is presented on a separate prompt sheet. The minimum number of word tokens per paragraphs is 25.

*Validation: 5% of the paragraphs could have less than 25 word tokens. 1% of the text C1_T can be missing in C1_PS.*

#### IV.2.3.2 C2_PS (news articles, novels and short stories with short sentences)

The prompt sheets are derived from C2_T. The complete text corpus C2_T must be used for C2_PS. The text is grouped into small paragraphs as described above. Nevertheless the paragraphs should be designed in favor to cover especially stress positions at the beginning and the end of the sentences. Each paragraph is presented on a separate prompt sheet.

*Validation: 1% of the text C2_T can be missing in C2_PS.*

#### IV.2.3.3 C3_PS (Domain: constructed phrases)

The prompt sheets are derived from C3_T. The complete text corpus C3_T must be used for C3_PS. Each phrase is a sentence. Each sentence is presented on a separate prompt sheet.

*Validation: 1% of the text C3_T can be missing in C3_PS.*

*IV.2.4 Corpus for the Pre-Selection of the Baseline Voices*

The corpus is denoted by C_PreR.

***Speaker selection for TTS database in RDI is done through the following procedure:-***

***Initial selection of speakers:-***

***To make sure that the mandatory requirements are satisfied. So this step results in initial rejection of some speakers.***

***In this step each candid speaker records a news sample of about 10 minutes. He is asked to read in newscaster speaking in neutral manner. Then a judgment committee consisting of 2-3 experienced linguists and 2 TTS researchers is constructed. This committee checks that each speaker meets the following specs:-***

1. ***Native with pure native dialect.***
2. ***Professional speaking style.***
3. ***Age ranging from 25 to 50.***
4. ***Very low falter and reading errors rate.***
5. ***Pleasant, consistent and uniform voice quality produced throughout the session.***
6. ***Accurate intonation, according to the sentence read***

***As these are mandatory specifications any speaker the committee decides he failed to pass any point of the preceding is rejected.***

***Final speaker selection:-***

***In this step recording of speakers who path initial step are then checked again rating each speaker in the following points:-***

1. ***Voice quality (Pleasantness, consistency) ,***
2. ***Pronunciation quality***
3. ***Speech correctness (low falter and reading errors rates)***
4. ***Intonation (accuracy and consistency)***
5. ***Pronounced phonemes boundaries easily distinguishable (using spectrogram).***

***The speaker with the maximum overall rate is selected.***

*IV.2.5 Baseline Corpus*

IV.2.5.1 Prompt Sheets

The prompt sheets are created from the complete corpus C_PS. The related prompt sheets are called
C1_BLPS (corpus is equivalent to C1_PS)
C2_BLPS (corpus is equivalent to C2_PS)
C3_BLPS (corpus is equivalent to C3_PS)

IV.2.5.2 Recordings (C_BLR)

The recordings are done under the conditions as described in chapter 6. The speaker reads the complete corpus C_PS of prompt sheets in a speaking mode as described in 3.3.1 and 3.3.2.

| Notation of Domain | Recorded Corpus |
|---|---|
| C1_BLR | Recording of read speech based on prompt sheets C1_BLPS |
| C2_BLR | Recording of read speech based on prompt sheets C2_BLPS |
| C3_BLR | Recording of read speech based on prompt sheets C3_BLPS |
| Status | Mandatory |
| Recommendation | |

| | |
|---|---|
| Comment | It is mandatory to read all prompt sheets; this could lead to deviations from the estimated amount of speech measured in h. |

*Validation: all the prompt sheets have to be read .1% of the word tokens read may be missing. The quality of the speech signals and signals from the laryngograph is checked according the validation criteria described in chapter 6.*

## IV.3 TTS Lexicon

For each language the TTS lexicon contains the pronunciation, the stress and syllable as well as POS information of words. Three lexica will be used and produced:

***In Arabic language the syllable structure is determined by order of consonants and vowels in a phrase. So for a word the syllable structure is meaningless is it will be affected by preceding and succeeding words. This will also affect stress locations.***

- common word lexicon as defined by LC-STAR for selection of speech segments
- extended common word lexicon containing all words of the corpus C_T, which are not included in the common word lexicon
- a small proper name lexicon.

The content of these lexica is specified in the next 2 sections.

### *IV.3.1 Common Word Lexicon*

The 'common word lexicon' is defined by the LC-STAR specifications (cf. Giulio Maltese et al. (2004) General and language-specific specification of contents of lexica in 13 languages. Version 2.1 (post-final)) [12]

Validation: according to validation criteria specified in the relevant LC-STAR deliverables.

### *IV.3.2 Extended Common Word Lexicon*

All words - except proper names - found in C_T and found in the transcriptions of C_R have to be included in the so-called Extended Common Word Lexicon in a format as specified by LC-STAR.

*Validation: The correctness of 500 phoneme transcriptions and 500 POS tags will be checked. 5% errors are allowed.*

### *IV.3.3 Small Proper Name Lexicon*

Within LC-STAR a large proper name lexicon is specified. Due to focus of TC-STAR on research and not on commercial applications such a proper name lexicon is not requested. Nevertheless names found in C_T have to be located in a 'Small Proper Name Lexicon' in a format as specified by LC-STAR.

*Validation: The correctness of 500 phoneme transcriptions will be checked. 5% errors are allowed.*

[12] http://www.lc-star.com/WP2_deliverable_D2_v2.1.doc; for Mandarin and Spanish already validated LC-STAR lexica exist.

# V Recording Environment and Recording Platforms

The usefulness of the recorded speech depends on the quality of the speech signal and on the precision with which the closure of the glottis can be marked reliably.
In section 6.1 and 6.2 the terms 'quality' and 'precision' together with their requirements are defined.
The sections 6.3 – 6.5 are proposals describing how to achieve the requested validation criteria.
In section 6.6 a procedure to perform the recordings is proposed.

## V.1 Quality of Speech Signal

The important parameters influencing quality are:

- signal to noise ratio ($SNR_A$) of the recorded speech
- linear phase distortion of the recorded speech
- reverberation of the room (measured by RT60)
- bandwidth of the speech signal

In Appendix A the measurement for $SNR_A$ and RT60 are defined. To minimize phase distortions a high sampling rate allowing using anti-aliasing filter with low linear phase distortion is requested. In order to achieve a high value for $SNR_A$ a high precision A/D-converter is recommended (24 Bit (optionally 16Bit) A/D converting precision).
Given these considerations following validation criteria have to be met:

- 96kHz sampling rate
- 24 Bit precision (16 Bit optional)
- $SNR_A >$ 40dBA
- RT60 < 0,3s [13]
- Bandwidth: at least 40Hz – 20 000Hz

Validation: the quality of the signals is checked according the procedure described in appendix A.

## V.2 Precision of Marking Epochs

The most important pitch event is the instant of glottal closure. This instant is called epoch[14]. For supporting the marking of the epochs automatically with the requested precision a laryngograph has to be used. To locate the pitch pulses precisely the reverberation of the room should be as low as possible and the signal of the laryngograph has to be suited. Nevertheless, finally only the precision of the marking is validated and it is up to the producer of the corpus how he achieves the precision required. In order to evaluate the precision of the pitch marking objectively a recording with laryngograph is mandatory.

- A synchronous signal of the laryngograph must be provided. This signal should have the quality[15], that the closure of the glottis can be derived reliable.

An extensive discussion concerning this issue is given in Appendix A. The validation criterium for the precision is described in section 7.3.

[13] A lower limit of 0,1s < RT60 is recommended in order to achieve a natural sounding voice. Recordings from anechoic chambers can be made natural in a post-processing step by applying reverberation algorithms.
[14] The name 'epoch' was proposed by Hartmut Pfitzinger (LMU, Munich)
[15] the quality of the signal of the laryngograph has to be defined

## V.3 Recording platform and devices

### *V.3.1 Recording plateform*

A platform with 2 synchronized channels:

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

is mandatory.
Optionally a stereo recording with the large membrane microphone can be done.
Further it is recommended to use a platform with 3 channels:

- channel 3: close-talk microphone

The close talk microphone is used to synchronize the speech signal with the signal from the laryngograph:
Caused by the movements of the speaker the distance between speaker's mouth/nostrils and the large membrane microphone can change by about 3 cm, giving a changing time shift of about 150 micro seconds between the speech signal of the microphone and the signal from the laryngograph. Under normal conditions these effects can be neglected[16] and the pitches of the speech recorded with the large membrane microphone can be detected directly with the signal of the laryngograph. (i.e. the close talk microphone is not needed). In extreme cases however jitter and Shimmer effects could result. For being prepared for this situation it is recommended to use a close talk microphone to compensate for the movements of the speaker.
All signals are sampled synchronously. It is recommended to store the signals directly to hard disc using an appropriate multi-channel recording hardware and software.

*Validation: synchronized signals recorded from a large microphone and a laryngograph has to be provided.*

### *V.3.2 Recording devices*

#### V.3.2.1 Large membrane microphone

This microphone is used to record the signal for the final voices. The distance to the speaker should be 60 cm or 30 cm with wind screen.

#### V.3.2.2 The Laryngograph

The laryngograph is needed to support the detection of pitch pulses (detect the start of the glottal closure). Experiences have shown that the laryngograph works not equally well for all voices. Furthermore it is critical how the laryngograph is mounted.

Validation: used for all recordings when not specified otherwise

#### V.3.2.3 Close-talk microphone

It is a head mounted microphone with a fixed distance of about 7 cm to the right of the mid-sagital plane at the height of the upper lip.

*Validation: can be used optionally*

[16] Due to investigation by H. Tillmann and H. Pfitzinger (Phonetic Institute of Munich) it can be concluded that those small varying delays are not relevant for marking pitch pulses for concatenative synthesis based on PSOLA principle.

### V.4 Recording procedure

One control person within the studio; one outside for technical control.
Prompting: from tilted TFT screen; speaker oriented in 32,8 degree angle to avoid reflection.
Recording unit each prompt sheet. Recording is repeated when an error occurs.
Recording in a short period to avoid quality changes if feasible.
Careful test of Laryngograph mounting (trying to find the optimal position).

### V.5 Segmentation and annotation

This section specifies the annotation of the speech database for the baseline voices.
For each utterance (speech file) it is required to provide:

- The prompt sheet used to elicit the utterance.
- The orthographic annotation.
- The (validated) phonetic transcription.
- A rough annotation of symbolic prosody.
- The segmentation into phones (can be performed partly automatically).
- The pitch marks, associated with the glottal closure (can be performed partly automatically).

| Feature | Prompt information |
|---|---|
| Value | For each utterance the prompt as presented to the speaker is provided. |
| Status | Mandatory |
| Recommendation | |
| Comment | |

### V.6 Transcription of the Recorded Speech

#### *V.6.1 Orthographic transcription*

| Feature | Orthographic annotation |
|---|---|
| Value | For baseline voices:<br>The text that was actually said by the speaker, transliterated. Furthermore, if the signal of a given word is not proper for concatenative speech synthesis, the word is preceded by the symbol '*'. Beginning and end of sentences should be marked. |
| Status | Mandatory |
| Recommendation | |
| Comment | In principle, the speech produced has to match the prompt. However, the orthographic annotation reflects what the speaker really said coping with minor deviations not detected during the recording phase.<br>The text is normalized using the standard procedure in speech synthesis: words are capitalized according to the lexicon; abbreviations are expanded into "normal" words, numbers and dates are transcribed, punctuation is detached from words, etc. The resulting text should remove ambiguities at the word level. The normalization scheme has to be documented in the language specific document.<br>The symbol '*' must precede any word which presents an evident |

| | |
|---|---|
| | problem for concatenative speech: noise (either from the speaker or external), mispronunciations, unintelligible words, word fragments, non-speech acoustic events, truncated waveforms, etc. |

*Validation: The orthographic transcriptions of 250 files (random sample from full db) are checked. A max. WER (Word Error Rate) of 0,1% is permitted. Ideally 5K words at least are validated.*

*V.6.2 Phonetic Transcription*

| Feature | Phonetic transcription |
|---|---|
| Value | For baseline voices:<br>The corpora are fully transcribed phonetically. The transcription has to be supervised to annotate what the speaker really said, including elision, reduction or assimilation present in continuous speech.<br>The phonetic transcription includes word and syllable boundaries and explicit mark of changes in the transcription caused by coarticulation between words (reduction, elision). Sentence beginning and end is also marked.<br>If in the orthographic transcription a word is tagged as 'problematic' then it doesn't need to be transcribed phonetically (however the word boundaries have to be marked and the symbol * is included instead of the phonetic transcription). |
| Status | Mandatory |
| Recommendation | |
| Comment | Each producing partner has to provide the used phone set. If the corpus contains foreign sounds which cannot be represented by the language phone set then additional symbols have to be included. Phones are limited by spaces, syllables by '-' and words by '\|'.<br>The 'pause' between words has to be included in the phone set and in the transcription. A pause is a silence with 'significant' duration. (>10 msec., before plosives, a perceived pause of length >100 msec.). The pause symbol is included before the word separation symbol. |

Validation: The phoneme transcriptions of 1000 phones are checked. A maximum of 5% PER is allowed.

*V.6.3 Prosodic Transcription*

| Feature | Symbolic prosody annotation |
|---|---|
| Value | For baseline voices and cross-language conversion voices:<br>- Phrase breaks are annotated using two levels: minor break (intermediate intonational phrase), major break (full intonational phrase).<br>- Pitch accent (intonational prominence) is annotated using two levels: 'normal' or 'emphatic' |
| Status | Mandatory |
| Recommendation | |
| Comment | The information about symbolic prosody complements the orthographic annotation: starting with the orthographic text, the symbol '#' is added in the words with pitch accent (## for emphatic words). The breaks are included between words as <b> (minor break) or <BB> major break.<br>Example: the #printer <b> run #out of #paper .<BB><br>Pitch marking should be based on the acoustic signal. |

*Validation: Pitch marking is validated taking into account the speech signal to check the correctness.*

*V.6.1 Segmentation*

| Feature | Phonetic segmentation |
|---|---|
| Value | For baseline voices:<br>All the baseline voices are segmented either automatically and/or manually.<br>The segmentation must match the manual phonetic transcription.<br>For each baseline voices at least two hours are manually segmented (or supervised).<br>For each phone, the starting and ending time must be provided. A 'middle' point can optionally be provided which indicates a reasonable point to split the speech segments in concatenative speech.<br>All the events (start and end positions) are indicated in seconds. |
| Status | Mandatory |
| Recommendation | It is recommended that the producers validate the automatic segmentation, reviewing those cases that may be problematic either for an error in the segmentation or in the signal itself: some cues can be derived from the alignment tool, the mismatch between phonologic characteristics and voice/unvoiced measures, abnormal durations, etc. |
| Comment | The phones which have being labeled as "problematic" do not need to be supervised. |

*Validation: The phoneme segmentations of 1000 phones are checked, 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.*

*V.6.2 Pitch Marking*

| Feature | Pitch marks |
|---|---|
| Value | Many systems use pitch synchronous processing. For this reason the baseline voices are labeled with pitch marks. The pitch marks are located at the instant of glottal closure as observed in the laryngograph channel.<br>The pitch marks are determined for all the baseline voices.<br>For the part of the baseline voices with supervised segmentation at least two hours are manually supervised. |
| Status | Mandatory |
| Recommendation | Although the use of laryngograph make pitch detection algorithms very reliable some errors can still happen, for instance when the signal is small. It is recommended that the producers validate the detection of the pitch marks, reviewing those cases that may be problematic either for an error in the detection or in the signal itself. Some cues are: value of the pitch, pitch derivative, mismatch between phonologic characteristic and f0 value, etc. |
| Comment | The phones which have being labeled as "problematic" do not need to be supervised.<br>The pitch marks are synchronized with the laryngograph channel. If |

| | |
|---|---|
| | needed the user of the database has to correct the delay between channels and adjust the instant to zero crossing. |

***Validation**: Deviation from ideal pitch mark is maximal 10% of duration of pitch period. and maximal 0,2 ms.*

## V.7 Database interchange format

This chapter defines the database interchange format for the NEMLAR TTS speech databases. It aims for compatibility with the existing SpeeCon databases and future TC-STAR TTS databases.

### *V.7.1 Storage Media*

Speech data will be stored on DVD's. An extra CD with only non-speech files (documentation, annotation, meta-files) will be considered.

### *V.7.2 File Types*

A complete database consists of signal and descriptive files. Signal files store the audio signals as recorded in the different recording scenarios. The descriptive files consist of annotation, documentation, and metadata files. The annotation – or *label* – files contain administrative data about the signal, and the various transcriptions of the audio signal; the administrative data can be collected automatically during recording, whereas the annotation is created manually by trained human annotators. The documentation files describe the database in sufficient detail, and the metadata files are created automatically and are used to perform formal checks for the completeness of the database.
The descriptive (text) files may be stored non-redundantly on one separate CD-ROM whereas all the signal files may be stored on DVDs containing only these signal files (names and structure as defined below), a copyright file (COPYRIGH.TXT, see 8.11), and a disk ID-file (DISK.ID, see 8.3).

Signal files
It has been agreed to adapt the ESPRIT Project SAM format and to store speech on data files containing only the signal waveform samples without any header. An associated ASCII label file will provide the annotation and transcription information.

It is assumed that a platform with at least 2 channels is used the third being optional:
channel 1: large membrane microphone
channel 2: laryngograph
channel 3: close-talk microphone (optional)

All signals are sampled synchronously with a sampling rate of 96kHz, 24 bit. (16 bit optionally)

### *V.7.3 Directory structure*

The directory structure is independent of the content of the speech files and thus allows a fully automatic creation of a file system during recordings. Documentation directories are added to the overall file system hierarchy during later processing.

Root directory and media name
The storage media for validation and distribution will be named according to the following scheme

```
<database><p><oo>
```

where <database> is defined in Table 8.2, <p> is one of "_", "D", or a digit "0"-"9", and <oo> a two digit code. For the code <p>, "_" is kept for SpeechDat compatibility, "D" is used for media containing only documentation data, and the digit may be used to denote any data disk with a sequence number higher than 99. The medium name is stored in file DISK.ID in the root directory. The following files will be present in the root directory of each DB:

| | |
|---|---|
| DISK.ID | 11-character disk identification <database><p><oo> where <p> is one of "_", "D", or a digit "0"-"9", and <oo> is a number from 00 to 99 |
| README.TXT | plain text database description file |
| COPYRIGH.TXT | plain text copyright file |

**Table 8.1 – Content of root directory**

README.TXT and COPYRIGH.TXT are formatted in the ISO-8859*n* character set, UTF 8 for Mandarin and other non-European countries respectively. README.TXT lists the contents in terms of files and file structures for the databases. DISK.ID is expected on every disk, including those with speech files. The numbering need not follow a continuous scale.

*Validation: Implementation of correct directory structure is checked.*

### *V.7.4 Speech and label file system hierarchy*

The general structure for all signal and the label files is

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

with "/" a generic file system separator symbol.

| | |
|---|---|
| <database> | Defined as <dbName><#><language code><br>where <dbName> is NMTTS, <#> is 7 for NEMLAR, <language code> is <ISO-639 code>: FOR OUR DATABASE: NMTTS7AR<br>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><br>Where <TTT> is the type of the subcorpus, i.e. either BLR (Baseline), |
| <subcode> | Defined as <TTT><SS><br>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 8.2 – Directory structure**

*Validation: Implementation of correct file name system is checked.*

### *V.7.5 Documentation directories*

The documentation will be held in a file system with the following structure

| | |
|---|---|
| / | 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 8.3 – Documentation files hierarchy**

*Validation: Implementation of correct documentation file structure is checked.*

### *V.7.6 File name conventions*

File names have to go beyond the subset of the ISO 9660 standard, i.e. file names will have more than 8 characters, viz. 11; they will have a 3 character file extension:

```
<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:<br>S: SAM label file<br>L: pitchmark file from laryngograph recordings (text)<br>P: phonetic segmentation file (text)<br>1,2,3: speech signal files for channels 1 – 3 |

**Table 8.4 – File name conventions**

The filename structure safeguards that each file has a unique name independent from the (sub)directory that it is placed in. This old SpeechDat principle should prevail over the ISO9660 8 character file name limitation.

*Validation: Implementation of correct use of file name conventions is checked.*

### *V.7.7 Speech file format*

All signals are sampled synchronously with a sampling rate of 96kHz, 24 bit (optionally 16 bit) with the least significant byte first ("lohi" or Intel format) as (signed) integers. A description of the sample rate, the quantization, and byte order used is held in the SAM label file.

### *V.7.8 SAM Label file format*

Given the need for some small modifications to the label formats, it was decided to introduce a new version number (version 6.1) for the modified SAM label files. Label files adhere to a modified SAM label format:

```
ABC: item_1, item_2,..., item_n
```

where

- ABC is a three letter mnemonic followed by a colon; the mnemonic must contain only 7-bit US-ASCII character and may not contain spaces or colons

- items after the mnemonic are separated by commas, i.e. they cannot contain commas themselves
- items can be empty
- spaces after the colon or in between items are recommended to improve readability
- a label line is delimited by <CR><LF>, the line end sequence according to the DOS operating system.

"A label file begins with the mnemonic "LHD:" and ends with "ELF:". The mnemonic "LBD:" splits a label file into two sections: the LABEL FILE HEADER and the LABEL FILE BODY. After LBD: only LBR:, LBO:, LBB:, LBP: and ELF: may follow."

There is one SAM label file assigned to each utterance (i.e. one for all the recording channels).

*Validation: 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.*

### *V.7.9 SAM Labels*

Contrary to SpeechDat/Speecon databases, in a TTS database, there is a single speaker that utters a lot of sentences. A single file containing the basic information of the speaker (sex, age, sampling rate,…) is therefore preferred and the annotation should include more information about segmentation, i.e. in LBB for each unit (BEG; END of utterance) instead of a sole broad phonetic transcription.
SAM label fields can be either

- free-form text,
- single items from a fixed vocabulary, or
- lists of attribute-value pairs.

The general principle is to allow as little freedom in filling in the label fields as possible to prevent editing errors, and to have meaningful label field entries that can be read by humans as well as machines. This means that mnemonic forms are used for items from a fixed vocabulary, e.g. for an indication of the acoustic environment. Whenever possible, attribute values should be binary, i.e. [ON|OFF]. If an attribute list is defined for a SAM label, then the SAM label field must contain all attributes with the appropriate values.

For the SAM files all data except for the transcriptions (LBO, LBP, LBB) is known at recording time. Hence the values for all SAM label fields can be automatically provided, with empty text parts for the transcription labels LBO, LBB, LBP. In Table 8.5 optional items are enclosed in braces "{}"; alternatives are enclosed in square brackets "[]" and separated by the vertical bar "|".

| **SAM Label** | **Description** | **Format** | **Format string** |
|---|---|---|---|
| LHD | Label header | Fixed vocabulary item: SAM 6.1 | %s |
| ELF | End of label file | | |
| CMT | Comment | Free-form text | %s |
| DBN | Database name | TC-STAR_TTS_ <LL> | %s |
| SCD | Speaker code | a 3-digit number | %03d |
| SEX | Speaker gender | Fixed vocabulary item: [M\|F] | %s |
| SNM | Speaker name | String | %s |
| AGE | Speaker age | Integer | %d |
| ACC | Speaker accent | Fixed vocabulary item from list of dialects | %s |
| PRF | Speaker profession | String | %s |

| | | | |
|---|---|---|---|
| NL1 | First native language | Character string with first native language of the speaker | %s |
| NL2 | Second native language | Character string with second native language of the speaker | %s |
| DIR | Speech file directory | Fixed vocabulary item from file system /<database>/<Type>/<sub code> | %s |
| SRC | Speech file names | A comma separated list of 11.3 file names | %11c.%3c, %11c.%3c, %11c.%3c,%11c.%3c |
| SCC | Scenario code | Database type, one of: BLR, VCR, ESR | %s |
| CCD | Corpus code | 2 character code, one of: 11, 12, 20, 31, 32, 33 | %2c |
| REP | Recording place | String representing the town of recording | %s |
| RED | Recording date | DD/Mon/YYYY | %02d/%3c/%4d |
| RET | Recording time | HH:MM:SS | %02d:%02d:%02d |
| BEG | Labeled sequence begin position | Integer | %d |
| END | Labeled sequence end position | Integer: number of sample points in recording – 1 | %d |
| SAM | Sampling frequency | Integer | %d |
| SNB | Number of (8-bit) bytes per sample | Integer: 3, (2), signed | %1d,%s |
| SBF | Sample byte order | Integer: [0\|lohi] | %s |
| SSB | Number of significant bits per sample | Integer: [24 / 16] | %d |
| QNT | Quantization | Fixed vocabulary item, e.g.: PCM | %s |
| NCH | Number of channels | Integer: 3 | %d |
| REV | Reverberation | Floating point | %f |
| SNQ | Signal/Noise Quality in dBA | Attribute value pair list,<br>CHN1 = %f, CHN3 = %f<br>The SNR values (dBA) of the microphone recordings | %f |
| BWI | Band width of speech signal | Typically 40-20000 Hz | %d-%d |
| LGG | Usefulness of Laryngograph signal | + or - | %c |
| LBD | Label file body | | |
| LBR | Prompt text | BEG,END,<gain>,<min>,<max>,<prompt text><br>with <gain>, <min>,<max> optional signal values; if they are not known, the values may be left empty, but the correct | %d, %d, %d, %d, %d, %s |

| | | | |
|---|---|---|---|
| | | number of commas must remain. <prompt text> is the<br>text that appears on the screen. | |
| LBO | Orthographic transcription | BEG,END,<sentences> | %d, %d, %d, %s |
| LBP | Prosodic transcription | BEG,END,<sentences> | %d, %d, %d, %s |
| LBB | Phonetic transcription | BEG,END,<sentences> | %d, %d, %d, %s |

**Table 8.5 – TC-STAR TTS SAM labels**

*V.7.10 Optional SAM labels*

In order to keep the annotation as concise as possible, only a few of the optional labels defined in SpeechDat(II) and SpeechDat-Car are also allowed in NEMLAR TTS.

| | Description | Format |
|---|---|---|
| TXF | Name of the prompt sheet text file | 8.3 file name |
| SYS | Labeling system | free form text |
| SPA | SAMPA version | free form text |
| ASS | Assessment code | free form text |

**Table 8.6 – Optional SAM labels**

*Validation: formal correctness of SAM label files is checked.*

*V.7.11 Other label files*

The other label files are the files with extensions:

- <LL>L: contains the time stamps of pitch markers from the laryngograph signal
- <LL>P: contains the phoneme transcriptions with time stamped segment boundaries.

Both files have to be provided in TXT format.

The time stamps are coded with one line for each epoch, with the time of the glottal closure expressed in seconds. The line ends with the sequence <CR><LF>.
The segment boundaries are coded with one line for each phoneme (including the symbol "pau" for pause). Each line consists on three fields separated by a tab stop. First, the phoneme symbol (using the SAMPA notation). Then, the starting time, and finally the ending time. All the times are expressed in seconds. The line ends with the <CR><LF> sequence.
The phonemes from this file must match the phonemes in the LBB field in the SAM label file. The only difference is that in the segmentation file the stress and boundary marks are not included.
Note that the precision of the time stamps should be enough to determine the sample with the recording sampling frequency. For instance, for 96kHz, 5 decimal positions are required.

*Validation: formal correctness of other label files is checked.*

*V.7.12 Table files*

The table files are mandatory database files providing an overview of the NEMLAR TTS database. They are created from the signal and/or label files of the database and formatted as follows:

- each record (= row) is delimited by the sequence <CR><LF> (ASCII 13 and 10)

- each field (= column) is delimited by a tab stop (`\t` in C, Java, perl; ASCII 9)
- numbers are written in their original format (both integer and real)
- dates are given in DD/Mon/YYYY with month names in English
- times are given in HH:MM:SS
- null fields are permitted and have no content ("null value" in DBMS terminology)
- field names are SAM labels, and they are given in the first line of the file

The table files are

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

The speaker and recording condition tables are related to each other. All data is stored in a DBMS-like structure, i.e. without redundancy and unique key values in each table. The relationship between tables is established by using a common SAM label in the related tables (in DBMS terminology the SAM labels are "attributes". A SAM label is a "primary key" attribute in one table, and in all related tables it is a "secondary" or "foreign key" attribute).
SPEAKER.TBL:
This file contains mandatory information about the speaker. To guarantee a unique identification key, speakers are given a speaker code SCD. This speaker code must be independent of the current recording session number so that it allowed recording the same speaker in more than one recording, but with the same SCD.
SPEAKER.TBL contains the following fields:

| | |
|---|---|
| SCD | Unique speaker code |
| SNM | Speaker name |
| SEX | Speaker gender |
| AGE | Speaker age |
| ACC | Speaker accent |
| PRF | Speaker profession |
| NL1 | First native language |
| NL2 | Second native language |

**Table 8.7 – SPEAKER.TBL fields**

REC_COND.TBL:
The recording condition table stores all information relevant to a recording session. It contains the following fields:

| | |
|---|---|
| SCC | Scenario code |
| CCD | Subcorpus code |
| SCD | Unique speaker code |
| REP | Recording place |
| RED | Recording date |
| RET | Recording time |
| TXF | Prompt sheet text file |

**Table 8.8 – REC_COND.TBL mandatory fields**

All fields are mandatory, except for TXF which is optional.

The lexicon file is an alphabetically ordered table of distinct lexical items which occur in the corpus with the corresponding pronunciation information. Each distinct word should have a separate entry. As the lexicon is derived from the database it must use the same alphabetic encoding for special and accented characters as used in the transcriptions. The lexicon table should be provided in the XML-format defined by LC-STAR [LC-STAR D3.0 ref] see http://www.lc-star.com/archive.htm

*Validation: formal correctness of table files is checked.*

### *V.7.13 Index files*

The index files allow quick access to speech and transcription data. The mandatory CONTENTS.LST file stores the transcription of the close talk microphone as given in the LBO, LBP, and LBB fields.
It is constructed from the SAM label files; in general, a Perl script or similar program creates this file. All attributes are mandatory, and empty attribute values must be left empty.

### *V.7.14 Contents index file*

The contents index file is a TAB delimited ASCII file that stores the transcription field (LBO) and relates it to properties of the signal file and speaker.

| | |
|---|---|
| DIR | directory |
| SRC | speech signal file name |
| SCC | scenario code |
| CCD | subcorpus code |
| SCD | speaker code |
| LBO | speech transcription without the numerical data |
| LBP | prosodic transcription without the numerical data |
| LBB | phonetic transcription without the numerical data |

**Table 8.9– Content index file definition**

*Validation: formal correctness of contents file is checked.*

### *V.7.15 Documentation files*

All files are mandatory except when they are explicitly marked optional.

#### V.7.15.1 Root directory

The root directory contains three mandatory files:

- COPYRIGH.TXT : a copyright text in ASCII format, mentioning also the NEMLAR project.
- DISK.ID : an 11-character string with the volume name (required for systems that cannot read the physical volume label),
- README.TXT : an ASCII text file that lists all files of the database, except for signal and label files which can be indicated by their name template.

An additional (XML or HTML) file, README.HTM, may optionally provide browser access to all documentation and selected signal and label files.

V.7.15.2 DOC directory

This directory contains documentation files, including a description of the database design and transcription manual in one of these formats:

| | |
|---|---|
| DOC | Microsoft Word text processor file |
| TXT | ISO 8859-1 DOS-formatted text file |
| PDF | Adobe Portable Document Format |
| PS | Adobe PostScript format |
| HTM | XML or HTML format |

**Table 8.10 – Content of the DOC directory**

All the documents in .DOC format have to be provided also in PDF or PS. In fact, if the document has not been produced using MSWord, then the DOC file is not required, but the PDF file has to be provided.

DESIGN.DOC
The DESIGN.DOC, in English, contains 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.

### Platform description

A complete description of the recording platform (in English) can be optionally provided as PLATFORM.DOC.

### Transcription manual

A complete transcription manual (in the native language with translation in English) can be optionally provided as TRANSCRP.DOC. This file, if provided, should hold the transcription

guidelines for orthographic transliteration, prosodic and phonetic transcriptions as well as segmentation and pitch labelling.

ISO8859 code table
It is mandatory that a sample character table corresponding to the current database is included in the database in PostScript format.

Phonetic Alphabet Definition Table
The SAMPA table used must be included in postscript format in the file SAMPALEX.PS. For languages that are not covered by SAMPA this file holds the phonetic alphabet definition. The lexicon information in DESIGN.DOC should provide a clear statement on which phonetic inventory is used.

Spelling variants
In many languages there are words or expression which can be spelled (i.e. written), in two or more different ways, e.g. "*all right*" vs. "*alright*" and "*colour*" vs. "*color*" in English, "*pra"*" vs. "*para a*" in Portuguese; these words are classified as heterographs. To maintain consistency, each site/language should compile a list of such items and include it on the CD-ROMs as **optional** file SPELLALT.DOC. The standard form must be before the alternate ones and it must be consistently used to transcribe what speakers said.
Finally, Table 8.12 gives an overview of documentation files:

| Directory | File | |
|---|---|---|
| TABLE | LEXICON.TBL | mandatory |
| | REC_COND.TBL | mandatory |
| | SPEAKER.TBL | mandatory |
| INDEX | CONTENTS.LST | mandatory |
| DOC | DESIGN.DOC | mandatory |
| | ISO8859<n>.PS | mandatory |
| | PLATFORM.DOC | optional |
| | SAMPALEX.PS<br>'INVENTORYLEX.PS' | mandatory |
| | SPELLALT.DOC | optional |
| | TRANSCRP.DOC | optional |
| | VALREP.{TXT\|DOC} | mandatory |

**Table 8.12 – Summary of Documentation files**

### *V.7.16 Recommendations*

*Data safety*
For data safety reasons it is strongly recommended to produce DVDs as soon as a sufficient number of sessions have been recorded or the hard disk is full, do not delete recording files unless two backup copies are safely stored in separate locations, e.g. on CD-ROM, format the removable hard disk prior to its (re-)use to detect bad blocks and to provide a clean and un-fragmented file system for the signal recordings.
*Storage*
It is advisable to make backup copies to safe media (e.g. DVD) as early as possible.

SAM label files
- It forbidden to split items over more than one line to facilitate processing.

Data processing
Use a relational database management system ("DBMS"), e.g. Microsoft Access or FileMaker to store the contents of the table files during processing:

- Data can be accessed quickly and consistently, and reports (including graphs) can be generated automatically
- Data is stored non-redundantly; hence, updates need to be made in one place only.
- For the final distribution, the relational database tables can be dumped to the distribution medium
- Tab stops are a natural field delimiter in DBMSs.

Document format
Note that tab stops normally are invisible chars on the screen and that some editors change them into spaces.

- Carefully format your documents according to the specifications in this document.

Note also that the appearance and printout of DOC files differ with the platform on which they are used; character encoding differs, and so do page counts.

- PS and PDF files must explicitly include the definitions for all fonts used in the document, including the standard PostScript set of fonts.
- Include PDF files of all word processor formatted files

## V.8 References

Ellbogen,T; Schiel, F., Steffen,A (2004):'The BITS Synthesis Corpus for German', Proc. LREC2004
Moreno, A. (2004) Specifications of lexicon interchange format. LC-STAR Technical report D3.0.
Kain,A.; Macon, M.W.: 'Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction' Proc. ICASSP 2001
Black, A.; Lenzo, K.: Building Synthetic Voices, FestVox 2.0 Edition ; http://www.festvox.org/
**Harald Höge, Antonio Bonafonte, Henk v.d. Heuvel, Asuncion Moreno, Herbert S. Tropf, David Sündermann, Ute Ziegenhain**. TC-STAR project Deliverable D8: (D8a_LIP) "Specifications of Language Resources for Speech Synthesis Language Independent Part (LIP)", V2.24 (Pre Final Version)

## V.9 Appendix A (from a document provided by Asuncion Moreno, UPC in the framework of TC-STAR) : Noise, Frequency Range, Reverberation and Recording

### *V.9.1 Frequency Range*[17]

The range of the speech signal should be 40 Hz – 20 kHz. Signals with frequencies outside of this range are caused by non speech. To remove the signals below 40Hz a suited high pass filter is needed. Such a filter could be constructed by filter with degree of 6 with an attenuation of 36 dB/octave. The attenuation should not be too large (maximal 48dB/Octave) to avoid distortions in the phase. In any case should be the filter linear in phase. Optimal is a Bessel filter.

### *V.9.2 Noise*

The noise on the speech signal should be as low as possible. There are several sources of noises:
- noise of the speaker
- background noise
- noise of the platform (amplifiers, anti-aliasing filter, A/D-converter, recording devices )

[17] notes of Guenther Ruske TU Munich http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/r/Ruske:G=uuml=nther.html

Breath noise of the speaker can be minimized by using a large membrane microphone in some distance. If the microphone is in the near field (30 cm distance) a wind screen should be used. Another option is to place the microphone in the fare field (60 cm distance).
Background noise can be measured by dBA or NC-xx. It is difficult to avoid noises at low frequencies. Good recording studios achieve a background noise level < 20dBA.
**The noise of the platform depends on the quality of shielding from electronic noise and the quality of the A/D converter. Due to the high sampling rate (96 kHz) and high precision of the A/D converter (24Bit per sample) the noise of the A/D-converter is minimal and no big requirements are needed for an anti-aliasing filter.**
For validating the quality of the speech signal only the speech signal is available. The final measurement will rely on the SNR achieved. This value is frequency dependent. A single value of SNR is given by $SNR_A$, as defined in B1.2.1. This definition includes the hearing characteristic of the human ear.

### *V.9.3 Definition and Measurement of $SNR_A$*

A-Weith is a standard for noise measurement that takes into consideration the human ear's sensitivity to certain frequencies (see Fletcher-Munson Curves). This is expressed as part of noise specifications and can be denoted by adding the letter 'A' to the spec - i.e. 15dBA.
Measurement:
Within a small frequency band with middle frequency f the energy of the noise $E_N(f)$ and the energy of the speech signal $E_S(f)$ is measured. This leads to
$dB_N(f)=Log_{10}(E_N(f))$
$dB_S(f)= Log_{10}(E_S(f))$
A weighted average with respect to f is calculated for $dB_N$ and $dB_S$ where the weights are defined according to the curve plotted in fig. B1.2.1 leading to $dBA_N$ and $dBA_S$[18]. Based on these values $SNR_A$ is defined as:
$SNR_A = dBA_S - dBA_N$

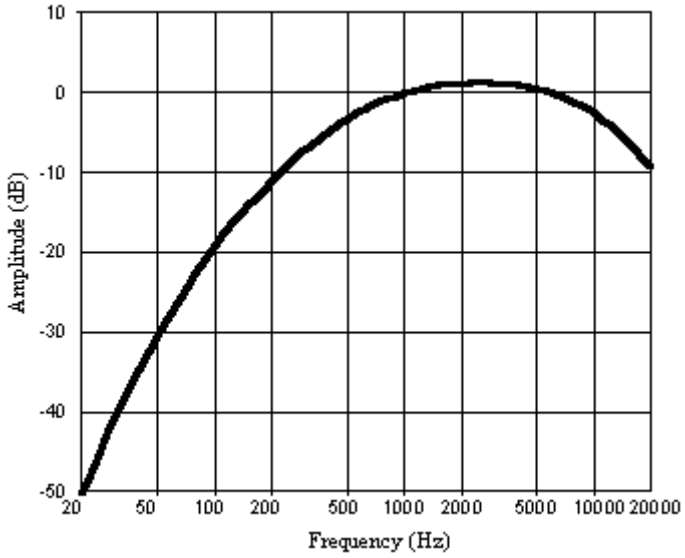


Fig B1.2.1 - Transfer function of the filter for measuring dBA.

[18] A matlab program calculating dBA by approximating the filter of Fig B1 is available at Siemens. This program has been made available kindly by the acoustic group (Prof. Hugo Fastl) of the Institute of man-machine communication at the technical university of Munich (TUM).

### *V.9.4 NC-xx*[19]

NC stands for Noise Criterion and refers to the quiescent or ambient background noise present in an acoustic space such as an auditorium or room. Curve, or contour, refers to the way in which our ears are sensitive to noise, which essentially follows the guidelines outlined by the Fletcher-Munson Curves, or other similar studies. In a nutshell this means that the human auditory system is not equally sensitive to noise at all frequencies. Further, as the noise level changes these relative sensitivities change with respect to one another. NC curves were developed to take all this into consideration, thus providing a reasonably objective way in which to document and communicate ambient noise levels in rooms. There are ratings given for various levels across the spectrum that take these curves into account. So a room with a certain amount of noise at 100 Hz will rate significantly better than a room with the same amount of noise at 1 kHz. Typical ratings range from NC-15 to NC-70. For example, a room said to meet the NC-15 requirement would be so quiet that the average listener would not perceive any background noise at all, yet there could be noise at 30 dB SPL below 80 Hz.
To determine the NC-xx level of a given room, all ambient noise present in the room is received by a microphone, and an octave or 1/3-octave filter bank is used to determine the noise energy within each band. A set of points is obtained from the pairs $\{f_c, E\}$, where $f_c$ are the central frequencies of the filters and E is the energy caught by them, in dB. In the next step, the lowest NC curve is selected so that all the energy points remain below it. The shapes of the existing NC curves, called NC-15, NC-20, NC-25, etc., are represented in figure B1.2.2, and are similar to the inverse transfer function of the filter for measuring dB A. As it can be seen, the noise is allowed to have more energy at low frequencies, where the human ear is less sensitive. For example, NC-15 or NC-25 are typical from quiet studios, while in a noisy environment like an office a noise level of NC-35 or NC-45 can be measured. A proposed instrument for the measurement of NC and RT is *TEF*, fabricated by *Techron* and sold by *Gold-Line* (USA).

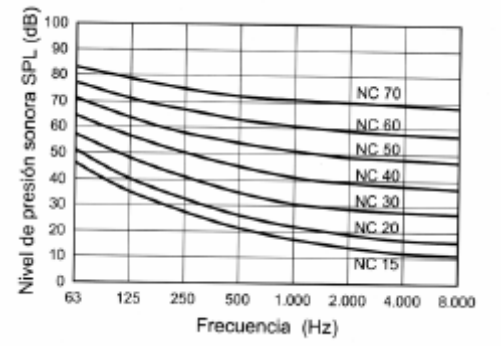


Figure B1.2.2 - NC curves

### *V.9.5 Reverberation RT-60*

Reverberation can be characterized by the reverberation time RT. The reverberation time of a room is the time it takes for sound to decay by 60 dB once the source of sound has stopped. Reverberation time is inversely related to sound absorption and is a way to measure the amount of absorption in a room. One procedure to measure RT60 consist of generating an excitation signal, and when the source has stopped the sound energy is measured during a time interval in each octave or 1/3 octave band. The RT60 of each band is calculated by determining the instant of 60dB decay. RT has a frequency depending value, because the absorption of a room is stronger at high frequencies, which wavelengths are short. Therefore, a reference value of RT is obtained at 1 KHz. When the ambient

[19] http://www.sweetwater.com/shop/studio/acoustic-treatment/glossary.php#55

noise of the room has a higher level than the 60dB decay threshold, the RT60 is indirectly measured by multiplying by a factor 2 the RT30. In good recording studios the value of RT60 is found to be less than 0.25 seconds. For a correct measurement, the microphone and the amplifier must be centered in the room, in a place with direct vision of all the surfaces. The minimum distance to all the surfaces must be greater than 1m. A proposed instrument for the measurement of NC and RT is *TEF*, fabricated by *Techron* and sold by *Gold-Line* (USA).

## V.10 Appendix B: Recording Equipments

### *V.10.1 Proposals for recording software*

- SpeechRecorder (C. Draxler, U. Munich)
- Samplitude 7.0 or higher.
- Steinberg Cubase or Nuendo (Imre Kiss; Nokia)
- NannyRecord (Antonio Bonafonte; UPC)

### *V.10.2 Proposals for recording Hardware*

RME HDSP 9652 Audio Card ca. 499 Euro
http://www.rme-audio.de/english/hdsp/hdsp9652.htm

RME OctaMic D (8 ch Preamp with ADAT output) ca. 999 Euro
http://www.rme-audio.de/english/micpreamps/octamic.htm
amplifier manual: www.rme-audio.de/english/download/octamic_e.pdf

### *V.10.3 Proposals for Large membrane condenser microphone*

- Neumann Type TLM 103 (600 – 1000€)
- Microtech Gefell M-930 (founded by Georg Neumann) (600 – 1000€)

### *V.10.4 Proposals for the Laryngograph*

- ***EG-2 Two-channel electroglottograph including 35mm electrodes***
- EG2-PC (2 channel electroglottograph with 35 mm electrodes; 20
- LaryngoGraph labatory version21;

### *V.10.5 Proposals for the Close-talk microphone*

- Cardioid condenser microphone (suppresses most surrounding noises):
- Sennheiser ME104
- Omni condenser microphone
- ShureWBH53B[22]

In order to synchronize the signal of the microphones and the laryngograph all signals have to be recorded in synchrony with time. Caused by the movements of the speaker, the distance between speakers mouth/nostrils and the large membrane microphone can change by about 3 cm centimeters,

[20] www.glottal.com;
[21] www.laryngograph.com; recommended by Bernd Möbius
[22] www.fullcompass.com/products/pages/SKU--65126/index.html

giving a changing time shift of about 150 micro seconds between the speech signal of the microphone and the signal from the laryngograph. Under normal conditions these effects can be neglected[23] and the pitches of the speech recorded with the large membrane microphone can be detected directly with the signal of the laryngograph. (i.e. the close talk microphone is not needed). In extreme cases jitter and Shimmer effects could result. For being prepared for this situation it is recommended to use a close talk microphone to compensate for the movements of the speaker

## V.11 Appendix C: MODERN STANDARD ARABIC Peculiarities

| | |
|---|---|
| Language: | Arabic (ara, AR) |
| Responsible NEMLAR partner: | ELDA & RDI |

This document describes the Arabic (Modern Standard) language specific peculiarities in the TTS database collection task of NEMLAR.

Arabic here refers to standard perceived pronunciation in the Arabic "formal" contexts and often referred to as Modern standard Arabic, mainly used in official speeches and the media. The most distinguishing features of Arabic are characters and vowel diacritics.

### *V.11.1 Description of character set used for orthographic transcription*

Unicode (Arabic scripts) is used for the orthographic transcription of Arabic

The character set used for Arabic (within Orientel) is ISO–8859-6.

### *V.11.2 Special handling of spelling*

No special handling of spelling.

### *V.11.3 Description of the Romanization scheme*

N/A

### *V.11.4 The Arabic Phonetic Alphabet*

see the page at: http://www.phon.ucl.ac.uk/home/sampa/arabic.htm

| Symbol | Keyword | English gloss | Orthography |
|---|---|---|---|
| | | **Consonants** | |
| Plosives | | | |
| b | ba:b | door | باب |
| t | tis?` | nine | تسع |

[23] Due to investigation by H. Tillmann and H. Pfitzinger (Phonetic Institute of Munich) it can be concluded that those small varying delays are not relevant for marking pitch pulses for concatenative synthesis based on PSOLA principle.

| | | | |
|---|---|---|---|
| d | da:r | home | دار |
| t` | t`a:bi?` | stamp | طابع |
| d` | d`arab | he hit | ضرب |
| k | kabi:r | large | كبير |
| g | gami:l | beautiful | جميل |
| | (in Egyptian pronunciation) | | |
| ? | ?akl | food | أكل |
| q | qalb | heart | قلب |
| p | paris | Paris | برس |

Fricatives

| | | | |
|---|---|---|---|
| f | fi:l | elephant | فيل |
| v | nivi:n | Nevien (personal name) | نفين |
| T | Tala:T | three | ثلاث |
| D | Dakar | male | ذكر |
| D` | D`ala:m | darkness | ظلام |
| s | sa?`i:d | happy | سعيد |
| z | zami:l | colleague | زميل |
| s` | s`aGi:r | small | صغير |
| S | Sams | sun | شمس |
| Z | Zami:l | beautiful | جميل |
| x | xit`a:b | letter | خطاب |
| G | Garb | west | غرب |
| X\ | X\ilm | dream | حلم |
| ?` (?\) | ?`alam | flag | علم |
| h | hawa:? | air | هواء |

**Nasals**

| | | | |
|---|---|---|---|
| m | ma:l | money | مال |
| n | nu:r | light | نور |

**Trill**

| | | | |
|---|---|---|---|
| r | rima:l | sand | رمال |

**Lateral**

| | | | |
|---|---|---|---|
| l | la: | no | لا |
| l` | ?al`l`ah | God | الله |

**Semivowels**

| | | | |
|---|---|---|---|
| w | wa:hid | one | واحد |
| j | jawm | day | يوم |

**Vowels**

| | | | |
|---|---|---|---|
| i | D`il | shadow | ظل |
| a | X\al | solution | حل |
| u | ?`umr | age | عمر |
| i: | ?`i:d | feast | عيد |
| a: | ma:l | money | مال |
| u: | fu:l | beans | فول |

*V.11.5 Some Frequently used Phrases from Orientel*

Details can be provided by ELDA.

- Digits
- Natural numbers
- Currency (Arabic common currencies: dinar, dirham, Jonaih, lira, and foreign currencies: Dollar, Euro, Pound Sterling, …)
- Yes/no
- Dates
- Weekdays
- Months (Islamic & Christians)
- Times
- Application words with transliterations and synonyms
- Telephone numbers
- City names
- Company names
- Personal names
- Foreign words