Technology and Corpora for Speech to Speech Translation
http://www.tc-star.org

Project no.:	FP6-506738
Project Acronym:	TC-STAR
Project Title:	Technology and Corpora for Speech to Speech Translation
Instrument:	Integrated Project
Thematic Priority:	IST

Deliverable no.: D8
Title: TTS Baselines and specifications

Due date of the deliverable:	M6
Actual submission date:	M12
Start date of the project:	1st of April 2004
Duration:	36 months
Lead contractor for this deliverable:	UPC
Authors	*Antonio Bonafonte (UPC), Harald Höge(Siemens AG), Herbert S. Tropf (Siemens AG), Asuncion Moreno (UPC), Henk van der Heuvel (SPEX), David Sündermann (UPC), Ute Ziegenhain (Siemens AG), Javier Pérez (UPC), Imre Kiss (Nokia)*

Revision: [final]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

1 Introduction

This document contains the specifications of Language Resources for speech synthesis, specifications for evaluation of speech synthesis systems and protocols to be applied between speech synthesis modules. In a speech synthesis system, the baseline is highly dependent of the language resources used, (normal speech, expressive speech, speaker) and performances can't be described independently of the speakers and styles used to train the system. For this reason the baselines descriptions and their performances will be described once the language resources have been collected and the baseline systems implemented.
This document is structured as follows, Section 2 contains the specifications of LR for Speech Synthesis, Section 3 describes the Specifications for Evaluation of Speech Synthesis Systems and Section 4 contains the protocols to be used between the different modules that form the Speech Synthesis System. Each section has their appendices and references

2 Specifications of LR for Speech Synthesis

2.1 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 for language resources (LR) based on which LRs in a variety of languages can be produced. The specifications will be developed within the framework of the EU- project TC-STAR (FP6-506738)[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 in long term 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.
During the production process of the LRs and also by using the LRs, changes and amendments in the specifications might become necessary in order to provide more adequate LRs[4]. For TC-STAR this document will be taken as the basis on which the LRs for the 3 languages mentioned above have to be created.
The section 2 of this document is one part of a deliverable which covers the language independent part (LIP) of the specifications of language resources. Language specific issues and language specific deviations from the language independent specifications are described in another TC-STAR document LSP (LSP denotes the Language Specific Part).
In case that peculiarities of a certain language make it necessary to deviate from the LIP specs, the LSP guidelines should be taken as the basis. Deviations should be properly explained and documented.
In the following sections of Chapter 2.1 the basic rationale behind the specifications is described. Chapters 2.2-2.9 focus on the specifications per se.

[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] This experience has been made during the process to specify LR for speech recognition (see specifications developed for the SpeechDat family: www.speechdat.org).

2.1.1 Focus of the section 2

This section describes the language independent specifications for language resources necessary for building speech synthesis systems and for investigating specific research topics in speech synthesis. In the context of TC-STAR the LR should be suitable for:

- building the most advanced state-of-the-art TTS systems. The TTS system built will also serve as a backend for a speech-to-speech translation system developed in this project.
- performing research on intra-lingual and cross-language voice conversion,
- performing research on expressive speech.

The creation of voices for TTS systems and research on voice conversion will be based on read speech**.** Text corpora are specified which have to be read by selected speakers. For research in expressive speech recorded data (e.g. recordings from the Spanish or European parliament) and read data will be used.

The main chapters of this section 2 are:

- the construction of the text corpora,
- the procedure to select suited speakers,
- the recording platform,
- the annotation of the recordings of the speakers,
- the database interchange format.

The language resources created according to the specifications will be validated. For validation specific validation criteria will be developed. Minimal requirements will be laid down in this document. The final validation criteria will be provided in a separate documentation.

2.1.2 Notation of Corpora

In the following paragraphs the corpora are denoted in general by Cn.m_xy , where

- n denotes the design principle of a certain scenario of the corpus C (i.e. 1: transcribed speech, 2: written text, 3: constructed phrases. If n is omitted the complete corpus is denoted).
- m denotes a certain sub-corpus in the given scenario (see section 2.1.3 for definitions). If m is omitted denotes the complete scenario n.
- x denotes the application of the corpus (e.g. BL: Baseline corpus, V corpus for voice conversion, EX for expressive speech); x is not always denoted.
- y denotes the content of the corpus: T (Text), PT (Prompt Text), R (Recorded speech) etc.; y is not always denoted.

2.1.3 Design Principles of the Text Corpora

The basic design principle relates to the term ‘suitability’. In the context of TC-STAR the term suitability refers to LRs which are optimally suited for generating most advanced state-of-the-art TTS systems covering different domains and for performing research on intra-lingual and cross-language voice conversion. Both aspects have to be considered in the design of the LRs. In this document however the design focuses on the first aspect, i.e. building the most advanced state-of-the-art TTS system. Considerable large effort is devoted to support research in voice conversion though.

For building a general purpose TTS system speech has to be synthesized for any given 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. The

[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’, and different domains as‘broad cast news’, ‘call home’ etc. are defined and related LRs are provided

document 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 different domains have been created. The domains selected in LC-STAR serve also 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' in various prosodic contexts are regarded. Throughout this document the term 'speech segment' refers mainly to triphones and syllables sometimes also to diphones taken as basic segments for speech synthesis. Although triphones or syllables are quite 'large' segments 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 besides from the domains covered within this project text from novels and a sub-corpus called 'frequent phrases' is specified which is constructed from domains as specified in LC-STAR.

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 are combined with certain prosodic events. Furthermore linguistic structures in written text (e.g. as found in a newspaper) differ from those found in spoken language. 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 of written text containing many short sentences (C2) is specified additionally.

With respect to the above mentioned considerations the text corpora are composed the following scenarios:

- C1: transcribed speech (transcribed speech from different domains)
- C2: novels and short stories with short sentences (written text from different domains)
- C3: constructed phrases (text specifically constructed)
- C4: expressive speech (transcribed speech from expressive speech)

Special corpora are needed for research in cross language voice conversion. For this issue parts of the corpus '*transcribed speech'* are translated into a given target language leading to parallel text corpora[7].

For this purpose C1 is split into 2 sub-corpora:

- C1.1: parallel transcribed speech (parallel text in 2 languages)
- C1.2: general transcribed speech

To achieve high coverage in various domains C3'*constructed phrases*' is composed by 3 sub-corpora:

- C3.1: 'frequent phrases': serves to improve the quality of frequently used phrases like phrases which contain dates, numbers, *yes/no* expressions and for frequent used phrases found in domains as defined in the LC-STAR specifications.

[6] www.LC-STAR.com

[7] Within TC-STAR a great part of the corpus 'transcribed speech' is derived from 'parliamentary speech'. This domain is used for the TC-STAR speech to speech translation demonstrator for the language pairs UK - ES and UK - Mandarin

- C3.2: ‘triphone coverage sentences’: serves to improve the coverage of speech segments with respect to missing or rarely seen triphones or syllables.
- C3.3: ‘mimic sentences’: This corpus serves for research in intra-lingual voice conversion. The corpus contains sentences with high coverage in all phonemes of a language including rare phonemes. The sentences have to be read in a ‘mimic mode’.

2.1.4 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 taken into account.
For building a single voice in a given language for a state of the art speech synthesis system a total volume of 10h of speech is considered to be adequate.
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 running words. This amount is distributed on the sub-corpora described above:

- **C1_T: transcribed speech (45 000 word tokens)**
  consists of:
  - C1.1_T: parallel transcribed speech (9 000 word tokens)
  - C1.2_T: general transcribed speech (36 000 word tokens)
- **C2_T: written text (27 000 word tokens)**
- **C3_T: constructed phrases (18 000 word tokens)**
  consists of:
  - C3.1_T: Frequent Phrases (8 000 word tokens)
  - C3.2_T: Triphone Coverage sentences (8 000 word tokens)
  - C3.3_T : Mimic Sentences (2 000 word tokens)
- **C4_T: expressive speech (9 000 word tokens)**

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 C1.2_T and C2_T however have to be derived from much larger corpora.

2.1.5 Building Voices and Related Recorded Corpora

Based on the text corpus C_T prompt texts are presented to the speakers on display resulting in the corpus C_PT . Using the prompt text different speakers are recorded to produce different ‘voices’.

2.1.5.1 Voices and Related Corpora for Building TTS-Systems

The voices recorded to build a TTS system (baseline system) are called ‘***baseline voices***’. For the baseline system 1 male and 1 female voice will be recorded in general. For each voice and for each language the duration measured in hours (h) will be approximately:

- C1_BLR: recorded transcribed speech (5h)
- C2_BLR: recorded written text (3h)
- C3_BLR: recorded constructed phrases (2h)

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

2.1.5.2 Voices and Related Corpora for Voice Conversion

For generating voices for research in cross-language voice conversion for each language pair 2 male and 2 female bilingual speakers are recorded. The speakers read the prompt text C1.1_PT derived from parallel transcribed speech in both languages. To generate the parallel transcribed speech transcribed text from the European Parliament speeches in English (as delivered by the commission) is selected and the original UK-English text is translated into Mandarin and Spanish. Translations can be adjusted in the target language (e.g. proper names, etc.). The speech recorded in two languages from a bilingual speaker is called a '***Cross–Language Conversion Voice'***.
For generating voices for research in intra-lingual voice conversion first the corpus C3.3_PT of a language is read by a 'template' speaker leading to a recorded corpus of a '***template voice'***. This voice has to be reproduced by '*mimic voices*' using a specific kind of speaking called 'Mimic Mode' mimicking the timing and the accentuation pattern but not pitch and voicing quality (cf. Kain2001). The speech recorded in the mimic mode in a language is called an '***Intra-lingual Conversion Voice***'.
In general, the '*mimic voices'* are recorded from the bilingual speakers and the '*template voice*' is recorded from one of the baseline speaker. The resulting corpus composed by the sub-corpora are:

- C1.1_VCR cross - language conversion voice corpus for a language
- C3.3_TPR template voice corpus for a language
- C3.3_MIR intra-lingual conversion voice corpus for a language

called the **'*Voice Conversion Corpus***' (**C_VCR**).

2.1.5.3 Voices and related Corpora for Expressive Speech

For generating voices for research in expressive speech in Spanish and English, 2 male and 2 female bilingual speakers (or 2 male and 2 female speakers per language) are recorded. For generating voices for research in expressive speech first the corpus C4_PT of a language is played by the 'original' parliamentary who pronounced it. The speaking stile, (rough intonation, expression, speed and pauses) has to be reproduced to generate the expressive voice. The speakers read the prompt text C4_PT derived from parallel transcribed speech in both languages. The speech recorded in two languages from a bilingual speaker is called an '*Expressive Voice*'.
The resulting corpus is called the **'*Expressive Voice Corpus***' (**C_EXR**).

2.1.6 Speaking Mode

Speakers should read the text in a manner that good results can be achieved concerning the quality of the speech synthesis system as well as the suitability for research. 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 the current state of art of concatenative speech synthesis the concatenation of speech segments selected from corpora with different speaking style and expressivity leads to unsatisfactory results[8]. Due to the restriction of the corpus to 10h speech recorded from one speaker it was decided to focus mainly on coverage of phonetic and prosodic variations. In the project the developed speech synthesis systems serve as backend for a Spoken Language Translation system. The voice should therefore sound as being uttered from a competent translator speaking in a rather neutral manner.
For research in expressive speech the voice style should sound as an original parliamentary speaker.

[8] If the concatenated speech segments are derived from speech sections with different speaking modes the synthetic speech sounds not 'consistent'. Personal communication from Nick Campbell (ECESS, Maribor June 2004).

2.1.7 Selection of the Speakers and Related Corpora

Base line speakers have to be selected very carefully. Selection criteria are pleasantness of voice and the suitability for speech synthesis based on concatenation and pitch synchronous manipulation. A specific procedure for the selection is defined in section 2.3.4.
For the selection of the bilingual speakers needed for cross-language voice conversion and expressive speech no specific procedure has been defined.

2.1.8 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:

- 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 a useful. However it has to be taken into account that not all persons deliver a useful signal from the laryngograph. This fact has to be given consideration when selecting the speakers.
To precisely locate the position of glottal closure the reverberation of the room should be as low as possible.
For 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.

2.1.9 Annotation

Annotation and segmentation is based on the following rules:

- All speech recordings are transliterated in normalized text form, 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 (except Mandarin) speech recordings are completely phonetically transcribed and manually checked listening to the real recordings.
- For baseline voices the speech recordings are completely segmented in speech segments on signal level. 2 h of speech are checked manually by the producer.
- For baseline voices the speech signal of the speech recordings is completely pitch marked. 2 h of speech are checked manually by the producer.
- For expressive speech voices the speech recordings are completely segmented in words on signal level.

2.1.10 Database interchange format

Two types of label files are used: 1. SAM files 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 categorization of domains in C3.3.

2.1.11 Validation Criteria

The minimum requirements for validation are integrated in the document. They are marked by:

'Validation: description of the validation criteria'

2.2 Languages

Within TC-STAR the following languages are covered:

- for the baseline system the languages are Spanish (Castilian) as spoken in Spain, English as spoken in England and Mandarin as spoken in China Mainland.
- research on cross-language voice conversion is done for the language pairs English-Spanish and English-Mandarin.
- research on intra-lingual voice conversion is done for the languages English, Spanish and Mandarin.
- research on expressive speech is done for Spanish.

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

2.3 Speakers and Speaking Modes

Speakers are needed to create:

- voices for the baseline system,
- voices to enable research in intra-lingual voice conversion,
- voices to enable research on cross-language voice conversion,
- voices to enable research on expressive speech.

Within this section 2 it is assumed that no specific speakers will be selected for intra-lingual voice conversion and expressive speech. The voices for intra-lingual voice conversion will be derived from speakers providing the cross-language and baseline voices.

2.3.1 Number of Speakers

The number of speakers for generating voices is specified per language except for cross-language conversion where it is specified per language pair. 'N*umber of speakers*' denotes

- for baseline voices the number of speakers providing the baseline voice,
- for cross-language conversion voices the number of speakers providing cross-language conversion voices,
- for template voices the number of speakers providing template voices,
- for mimic voices the number of speakers providing mimic voices.

For specific languages the '*number of speakers*' could be less than specified[9]. This will be documented in the LSP.

Number of speakers	Kind of voice
1	Baseline voice male
1	Baseline voice female
2	Cross-language conversion voice male
2	Cross-language conversion voice female
1	Template voice[10] (template speaker)
2	Mimic male voice[11]
2	Mimic female voice[12]

[9] e.g. For Mandarin we will have only one single baseline voice; either male or female.

[10] This voice can be produced by one of the baseline speakers. Beforehand however it has to be controlled if the bilingual speakers are able to mimic the baseline speaker's prosody. Otherwise another suited template speaker has to be found.

[11] For this voice the bilingual male speakers can be used.

[12] For this voice the bilingual female speakers can be used.

2	Expressive speech male voice
2	Expressive speech female voice
Status	Mandatory if not specified otherwise in the LSP document
Recommendation	None
Comment	None

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

2.3.2 Speaker Profile

Feature	Native / bilingual speakers
Native skills	**For baseline voices:** 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. **For cross-language conversion voices:** The speaker for a given language pair should ideally be a bilingual speaker of these languages. If no adequate bilingual speakers are available with respect to the languages English, Spanish and Mandarin chosen in TC-STAR, also speakers should be accepted whose mother tongue is Spanish or Mandarin (depending on the language combination), respectively, and who speak English in an almost native way. **For expressive speech:** Ideally the speakers should be bilingual speakers of these languages with ability to roughly imitate speaking styles. If speakers with these abilities can not be found, 2 male speakers and 2 female speakers from each language with ability to roughly imitate speaking styles are preferred.
Status	Mandatory
Recommendation	For baseline voices: Both parents of a speaker should be fluent speakers of the given language as well.
Comment	none

Feature	Age of speakers
Value	22 - 50
Status	mandatory
Recommendation	none
Comment	none

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	none

Validation: During the speaker selection process the speaker's profile will be assessed (see 3.4).

2.3.3 Speaking Modes

2.3.3.1 Speaking style

Feature	Speaking style
Value	**For baseline and voice conversion:** Fluent reading compatible with a professional translator speaking in neutral manner. **For expressive speech:** Ability to roughly imitate speaking styles.
Status	Mandatory
Recommendation	None
Comment	No (imitated) spontaneous speech, no dialogue style.

2.3.3.2 Voice Quality

Feature	Voice quality
Value	Pleasant; consistent, even, uniform voice quality produced by each speaker throughout all sessions.
Status	Mandatory
Recommendation	None
Comment	None

2.3.3.3 Expressivity

Feature	Expressivity
Value	**For baseline and voice conversion:** The dominant expressivity is that chosen by the speaker compatible with a professional translator speaking in neutral manner. **For expressive speech:** Compatible with parliamentary sessions.
Status	Mandatory
Recommendation	None
Comment	None

2.3.3.4 Mimic Mode

Feature	Mimic Mode
Value	For intra-lingual voice conversion speakers of the same language try to mimic the timing, the accentuation pattern and the pitch contour of given template utterances. In order to assure that the speakers feel comfortable mimicking the given pitch contour the pitch of the template speech is varied adding an offset by means of PSOLA. This calibration is executed for each considered speaker. The procedure for calibration is described in Appendix A
Status	Mandatory
Recommendation	None
Comment	None

Validation: During the speaker selection process the speaker's speaking mode will be assessed (see 2.3.4).

2.3.4 Casting of speakers

2.3.4.1 Selection of Speakers for the Baseline Voices

Pre-selection phase:

Step 1:

Per given language and gender, a group of 5 speakers is selected which complies with the requirements of the speaker profile as defined in section 2. 3.2.

Step 2:

In order to spot pleasant voices, to detect obvious and non-obvious speech defects[13], to test the skill for reading text in an adequate style and to test the quality of the signal from a laryngograph, examples from 5 speakers are recorded. Two kinds of recordings are made:

The first recording should contain about 5 min. of speech recorded during the first interview made with the speaker.

The second recording is made under the recording conditions as described in chapter 2.6. The speaker has to read selected prompt text from C_PT covering all domains C1, C2, C3. The amount of speech recorded should be about 5 minutes.

Step 3:

3 experts (2 experts for synthesis, 1 phonetician/logopaedician being native speakers of the language in question) inspect the recordings of the given group of speakers (cf. step 2) including the signal delivered from the laryngograph and select 2 speakers for final selection. The final selection should be based on following criteria:

- affliction with speech defects,
- voice quality after F0 manipulation using pitch synchronously labeled speech units,
- pleasantness of voice,
- suitability of the signal from the laryngograph for pitch marking.

Final selection phase:

Step 4:

For each of the 2 selected speakers 200 different sentences will be recorded as described in chapter 2. 6. From these recordings 10 sentences based on concatenated diphone synthesis are generated (for a detailed description see section 4.6).

Step 5:

The 3 experts mentioned above inspect 10 sentences of each of the 2 speakers and select the one based on the following criteria:

- pleasantness of voice,
- voice quality based on the synthesized sentences.

***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 can be 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 can be submitted to the validation centre. Here the validation centre has only an advisory role.*

[13] Typical speech defects even among trained speakers might be related to *s*-, *th*-, *r*- or other sounds, missing distinction between voiced/unvoiced, incorrect nasalization or aspiration, etc..

2.3.4.2 Selection of the Speakers for Cross-Language, Conversion Voices and Expressive Speech

The selection process is described in the LSP. At least 5 min of speech is recorded for each potential candidate. The recordings are performed in an environment as described in Chapter 2.6 using a laryngograph.

Validation: the selected speakers will be assessed by auditory inspection on native tongue, age and proficiency (by trained native speakers of the languages). A short speech sample of each selected speaker can be submitted to the validation centre. Here the validation centre has only an advisory role.

2.3.4.3 Selection of the Speakers for Intra-Lingual Conversion Voices

No additional speakers for intra-lingual voice conversion are selected.

2.3.4.4 Selection of Template Speaker

In general, one of the baseline speakers is selected as Template Speaker.

2.4 Specification of Corpora

2.4.1 Amount of Corpora

Within this section 2 in total 6 corpora are specified:

- corpus for the pre-selection of the baseline voices
- corpus for the pre-selection of the conversion voices
- corpus for the final selection of the baseline voices
- corpus for the creation of baseline voices
- corpus for the creation of conversion voices
- corpus for the creation of expressive speech.

With the exception of the sub-corpus of recordings of the interview during the selection phase of the speakers 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 2.4.2.

2.4.2 Kind and Size of Sub-corpora of Corpus C_T

The text corpus C_T is composed out of the sub-corpora C1.1_T, C1.2_T, C2_T, C3.1_T, C3.2_T, C3.3_T and C4_T as documented in the table below.

The sub-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 parallel text corpus of transcribed speeches[14] from 2 languages. Corpus C2_T consists of text chosen from novels and short stories. 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 coverage on speech segments including rare speech segments. If not otherwise noted the size of the sub-corpora given by the number of word tokens is defined per language. The text corpus C4_T is a parallel text corpus of transcribed speeches from the parliament. Additionally the original recordings are also available.

Notation of Text	Kind and Size of Sub-corpora of Corpus C

[14] For TC-STAR C1.1_T should be transcribed '*parliamentary speeches'* translated from English to Spanish and Mandarin.

Corpus		
C1_T consists of:	**Transcribed speech** [15]	**45 000 word tokens**
C1.1_T	Parallel transcribed speech in 2 languages	9 000 word tokens per language
C1.2_T	General transcribed speech	36 000 word tokens
C2_T	**Novels and short stories with short sentences**	**27 000 word tokens**
C3_T consists of:	**Constructed Phrases**	**18 000 word tokens**
C3.1_T	Frequent used phrases	8 000 word tokens
C3.2_T	Triphone coverage sentences	8 000 word tokens
C3.3_T	Mimic sentences	2 000 word tokens
C4_T	**Expressive speech**	**9 000 word tokens**
Status	mandatory	
Recommendation	None	
Comment	None	

Validation: For C_T the correct minimum number of word tokens per kind of scenario and subcorpus will be checked from the prompt text. The number of word tokens can deviate by a maximum of 5% per scenario.

2.4.2.1 Domains and Size of Corpus C3.1_T 'Frequent Phrases'

The domains D1 – D5 chosen were also used in LC-STAR to define the domains for the common word lexicon. From these domains frequent used phrases have to be designed as derived from written text (article, news paper, etc.). D0 is included to cover the most frequent expressions relevant for typical TTS applications. The related phrases should be designed as utterances typically used in 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 LSP based on the following recommendations:

Domains	**Sub-Domains**	**Further descriptions**	**Size in %**
D0. frequently used colloquial expressions	D0.1 numbers (cardinal and ordinal)[16]	0,.., 10, tenth, hundred, ..., billion and all numbers which cannot be derived by concatenation rules (e.g. eleven, twelve, first, second,...)	50%
	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, ,years around 2005, special days	2004-12-29; 6 pm; 1996; Christmas; Monday; March	
	D0.4 seasons; time expressions	Spring; the day after tomorrow	
	D0.5 abbreviations	.com, .info, .net, .info, .org; EU; EC; TC-STAR; IBM;	

[15] The specific domains selected are defined in the LSP.

[16] 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 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	
	D0.6 miscellaneous	Sorry, you are welcomed, thanks,	
D1. Sports/Games	D1.1.Sports (special events)	soccer, skating, skiing, tennis, baseball, betting, lottery etc.	10%
D2. News	D2.1. Local and international affairs	top stories on domestic and foreign affairs, headlines with articles, etc.	10%
	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.	10%
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.	10%
	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	10%
	D5.2. Popular science	articles for lay people	
	D5.3. Consumer technology	Descriptions & manuals of mobile phones, PDA's, TV, video recorder, etc.	

Validation: distribution (percentage) of different domains could be checked.

2.4.3 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 sub-corpora C1.2_T, C2_T, and C3.2_T. The other text sub-corpora C1.1_T, C3.1_T, C3.3_T and C4_T are designed according to other criteria.

2.4.3.1 Definition of Speech Segments

In this document the speech segments of a given language are either triphones or syllables, except otherwise documented[17]. The set of speech segments used is documented in LSP. Each speech 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,

[17] E.g. for Mandarin

- End of a prosodic phrase.

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

- End-statement
- End-question
- 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.[18]

Prosodic properties for tonal languages

A more detailed definition of the prosodic properties on position, stress and for tonal languages can be found in LSP.

2.4.3.2 Specification of Segment Coverage Symbolic Coverage

The speech segments found in the baseline text corpus should cover a high percentage of the speech segments (stressed and unstressed for languages with stress patterns) found in the LC-STAR common word lexicon (counted on the LC-STAR lexicon discarding triphone singletons). The value achieved for speech segment coverage is documented in the LSP.
A coverage of 95% is recommended. The minimum requirement is 90%.

Prosodic coverage on position

Not all speech segments will be found in all positions and in all stress modes. However more frequent speech segments should have a higher coverage than less frequent ones. As a proposal to define coverage the LC-STAR wordlists (available with frequencies) could be used as a basis and than the following scheme could be developed:
X% of the most frequent speech segments should cover certain positions,
Y% of the most frequent speech segments which could be stressed should be covered by both: stressed and unstressed for languages with stress patterns.
Provided that the constraints on triphone coverage are too demanding it is also accepted to work with diphones (or longer segments) for computing the prosodic coverage. In the LSP a threshold for X and Y should be specified which is used to select significant segments to be covered in different positions. For each selected segment the number of categories that are used to characterizing prosody should be specified and well documented.

Tone coverage on position

The coverage of syllables with different tones in tonal languages should achieve 95%. The minimum requirement is 90%.

Approaches will be defined in the LSP.

Validation: the documentation will be checked.

2.4.3.3 Achievement of Segment Coverage

The method to construct the corpora is not mandatory. It is a 'best practice' proposal.

[18] Lexical stress does not exits in languages like Mandarin; this issue will be addressed in LSP.

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 1 000 000 word tokens
C2_T	novels and short stories with short sentences 1 000 000 word tokens
Status	Mandatory
Recommendation	None
Comment	None

Some algorithms to achieve high coverage are documented in appendix A.

2.4.4 Prompt Texts C_PT

The prompt texts are derived from the text corpora Cn.m_T. All prompted text is equivalent to the text corpora described in the previous chapter. The only difference is a notational one: prompt texts denote the text which will be displayed to the speakers.
The collection of all prompted text is called C_PT.

Validation: cf. Section 2.4.2

C1_PT (Transcribed speech)
The prompt text C1_PT is derived from the complete C1_T. The text sub-corpora C1.1_T and prompt text sub-corpora C1.1_PT consist of 2 languages. 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 displayed on separate prompt texts. The minimum number of word tokens per paragraphs is 25.

Validation: 5% of the paragraphs could have less than 25 word tokens.

C2_PT (novels and short stories with short sentences)
The prompt text C2_PT is derived from the complete C2_T. The text is grouped into small paragraphs as described in section 4.4.1. 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 displayed on a separate prompt text.

C3_PT (Domain: constructed phrases)
The prompt text is derived from the complete C3_T.

C4_PT (Domain: expressive speech)
The prompt text is derived from the complete C4_T.

2.4.5 Corpus for the Pre-Selection of the Baseline Voices

The corpus is denoted by C_PreR.

Feature	Quantity and Sort of Recordings
Value	5 min clean dialog speech by the speaker about his personal/professional

	background. 5 min read speech; recordings are made under the conditions as described in chapter 2. 6. The speaker has to read selected prompt sheets from C_PT covering all domains C1, C2, C3. The signal of a laryngograph is recorded synchronously.
Status	mandatory
Recommendation	None
Comment	This database serves also to check the recording platform.

Feature	Phonetic and Phonetic Coverage ; Annotations
Value	No requirements

Validation: The speech quality of the read speech and the signals from the laryngograph are checked according to the validation criteria described in chapter 2. 6.

2.4.6 Corpus for the Final Selection of the Baseline Voices

2.4.6.1 200 sentences corpus (C_200SR)

Feature	Quantity and Kind of Recordings
Value	Prompt text is selected from C_PT. The prompt text should cover at least 200 sentences leading to recorded speech of about 20 min of speech. The recordings are made under the conditions as described in chapter 2. 6. The signal of a laryngograph is recorded synchronously.
Status	mandatory
Recommendation	(cf. Phonetic coverage; annotation below).
Comment	None

Feature	Phonetic Coverage; Annotation
Value	The 200 sentence text corpus should be suited to synthesize 10 new sentences using diphone technology; i.e. the diphones found in the text corpus should be sufficient to construct 10 new sentences. The 10 new sentences have to be different to the original 200 sentences. The most equal sentence compared to a new sentence should have a maximum on overlap counted in words of 70%. The recorded speech must be annotated and segmented as described in chapter 2. 7.
Status	mandatory
Recommendation	
Comment	Additionally this database serves to check the annotations.

Validation: the corpus must contain 200 sentences. The recorded speech must be annotated and segmented as described in chapter 2.7.

Out of the recorded and annotated 200 sentence corpus 10 new sentences for testing the acceptability of potential concatenative synthesis using this voice are constructed leading to the 10 Sentence Corpus (C_10SR) for Testing Diphone Synthesis Acceptability.
The test scenario is presuming the example of a diphone approach and TD-PSOLA-based concatenation.

Validation: the corpus must contain 10 sentences. Each sentence or phrase should have a maximal word overlap of 70% to each of the 200 sentences and should contain at least 5 syllables.

2.4.7 Corpus for the Selection of the Conversion Voices and Expressive speech voices (C_5MR)

Feature	Quantity and sort of text
Value	At least 5 min of speech is recorded for each potential candidate. The recordings are performed in an environment as described in chapter 2. 6 using a laryngograph.

Validation: the minimum length of 5 min is checked.

2.4.8 Baseline Corpus

Recordings (C_BLR)

The recordings are done under the conditions as described in chapter 2.6. The speaker reads the complete corpus C_PT 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 C1_PT
C2_BLR	Recording of read speech based on C2_PT
C3_BLR	Recording of read speech based on C3_PT
Status	Mandatory
Recommendation	
Comment	It is mandatory to read all prompt text. Note: this could lead to deviations from the estimated amount of speech measured in h.

Validation: cf. Section 2.4.2. *The quality of the speech signals and signals from the laryngograph are checked according the validation criteria described in chapter 2.6 and Annex B.*

2.4.9 Cross-language Voice Conversion Corpus

The cross language voice conversion corpus is build from C1.1_PT (prompt text of parallel corpus of transcribed speech).
The bilingual speaker reads the parallel texts in the languages given. Recording is done according to the recording conditions as described in chapter 2. 6.
The corpus is denoted as C1.1_VCR.

Feature	Recorded Corpus
Value	Recordings of read speech in 2 languages based on C1.1_PT
Status	Mandatory
Recommendation	None
Comment	None

Validation: cf. Section 2.4.2. *The quality of the speech signals and signals from the laryngograph are checked according the validation criteria described in Chapter 2.6 and Appendix B.*

2.4.10 Intra-Lingual Voice Conversion Corpus

The prompt text is C3.3_PT. It has to be read by the bilingual speakers in a mimic mode (see section 2.3.3.4). They mimic a template voice. This template voice is derived from a modified

version of C3.3_BLR generated from the baseline speaker. The modifications are specified in Appendix A.

Feature	Recorded Corpus
Value	Recordings of read speech from C3.3_PT (2 000 word tokens)
Status	mandatory
Recommendation	None
Comment	The mode of reading is by the mimic mode.

Validation: cf. Section 2.4.2. *The quality of the speech signals and signals from the laryngograph are checked according the validation criteria described in Chapter 2 6 and Appendix B.*

2.4.11 Corpus for expressive speech

The prompt text is C4_PT. It has to be read by the bilingual speakers after listening to what the original speaker had said trying to imitate his/her style.

Feature	Recorded Corpus
Value	Recordings of read speech from C4_PT (9 000 word tokens)
Status	mandatory
Recommendation	None
Comment	The mode of reading is by imitating speaking style.

Validation: cf. Section 2.4.2. *The quality of the speech signals and signals from the laryngograph are checked according the validation criteria described in chapter 2.6 and Appendix B.*

2.5 TTS Lexicon

For each language the TTS lexicon contains the pronunciation, information about stressed syllables or tone markers for and syllable boundaries as well as POS information. The following lexica will be used and produced:

- common word lexicon containing all words of the corpus C_T, at prompt level and orthographic transcription level with POS and lemma information, and phonetic transcription as defined in LC-STAR,
- a small proper name lexicon.

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

2.5.1 Common Word Lexicon

The common word lexicon consists of a lexicon (see above) including at least all words - except proper names - found in C_T and found in the transcriptions of the C_R in a format as specified by LC-STAR specifications (cf. Giulio Maltese et al. (2004) General and language-specific specification of contents of lexica in 13 languages. Version 2.1) [19]

Validation: The correctness of the phonemic transcriptions and POS tags from the lexicon will be checked. 5% errors are allowed.

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

2.5.2 Proper Name Lexicon

All proper 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 phonemic transcriptions will be checked. 5% errors are allowed.

2.6 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 2.6.1 and 2.6.2 the terms 'quality' and 'precision' together with their requirements are defined.
The sections 2.6.3 and 2.6.4 specifies requirements for the recording platform and recording devices. Some suggestions of providers of platforms, recording SW and hardware and recording devices are given in Appendix B

2.6.1 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 B the measurement for SNR_A and RT60 are defined. To minimize phase distortions a high sampling rate allowing to use 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 ,[20]
- Bandwidth: at least 40Hz – 20 000Hz.

No post filtering is accepted to deliver the database.

Validation:

- *SNR_A > 40dBA must be achieved for 90% of the speech; SNR_A measured on labeled data,*
- *Clipping less than 0.1%,*
- *Digitizing: 24bit A/D accuracy(16 bits optional), 96KHz sampling),*
- *Frequency range 40 - 20 000Hz; 0.5dB deviation (channel after the microphone has flat frequency response in this range). The frequency range and frequency response of the acquisition system should be documented.*
- *Reverberation has to be in the range** : RT60 < 0.3s; has to be documented for a typical session.*
- *Use of a Laryngograph,*
- *Optional close talk microphone & stereo recording.*

[20] 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.

2.6.2 Precision of Marking Epochs

The most important pitch event is the instant of the glottal closure. This instant is called '*epoch*'. A laryngograph has to be used to automatically support the marking of the epochs for achieving the requested precision. It is recommended that to locate the pitch pulses precisely the reverberation of the room should be as low as possible and the signal of the laryngograph should 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 the recording with a laryngograph is mandatory.

- A synchronous signal of the laryngograph must be provided. This signal should have such a quality[21], that the closure of the glottis can be derived reliable.

The validation criteria for precision are described in section 2.7.3. and Annex B.

2.6.3 Recording platform

A platform with 2 synchronized channels is mandatory:

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

Optionally a stereo recording with the large membrane microphone can be performed.
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[22] 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 additionally 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. This could be checked in the documentation.

2.6.4 Recording Devices

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.

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.

[21] The quality of the signal of the laryngograph has to be defined.

[22] Personal communication by H. Tillmann and H. Pfitzinger (Phonetic Institute of Munich): from their investigations it can be concluded that these small varying delays are not relevant for marking pitch pulses for concatenative synthesis based on PSOLA principle.

Validation: It is checked if there is a laryngograph file for each speech file and vice versa (unless specified otherwise).

Close-talk microphone

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

2.6.5 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 text. 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).

2.7 Segmentation and annotation

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

- the prompt text used to elicit the utterance,
- the orthographic annotation,
- the phonetic transcription,
- a rough annotation of symbolic prosody,
- the segmentation into phonemes (can be performed partly automatically); for expressive speech it can be optionally provided.
- the pitch marks, associated with the glottal closure (can be performed partly automatically).

Feature	Prompt text information
Value	For each utterance the prompt text as presented to the speaker is provided.
Status	Mandatory
Recommendation	None
Comment	None

2.7.1 Transcription of the Recorded Speech

2.7.1.1 Orthographic transcription

Feature	Orthographic annotation
Value	For baseline voices and conversion voices: A transliteration of what was actually said by the speaker. Furthermore, if the signal of a given word is not suited for concatenative speech synthesis the word is preceded by the symbol '*'. Punctuation marks from the C_PT should be preserved. For expressive speech voices: A transliteration of what was actually said by the speaker. Furthermore, if the signal of a given word is not proper for concatenative speech synthesis, the word is preceded by the symbol '*'. Filled pauses such as *uh*, *um, er, ah, mm,* will be transcribed as well and the transcription will be between brackets []. Punctuation marks from the C_PT should be preserved.

Status	Mandatory
Recommendation	None
Comment	In principle the speech produced has to match the prompt text. However, the orthographic transliteration reflects what the speaker actually said coping with minor deviations not detected during the recording phase. The text is normalized using the standard procedure in speech synthesis: words are capitalized according to the lexicon; abbreviations are expanded into "normal" words as being pronounced by the speaker, numbers and dates are transcribed, punctuation is detached from words, etc. The resulting text should remove ambiguities at the word level. 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. For the baseline voices and conversion voices filled pauses are not expected. For expressive speech they should be properly marked. The normalization scheme and guidelines for transliteration have to be documented in the language specific document (LSP).

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.

2.7.1.2 Phonetic Transcription

Feature	Phonetic transcription
Value	For baseline voices and conversion voices: The corpora are fully transcribed phonetically. The transcription has to be 100% supervised to annotate what the speaker really said, including elision, reduction or assimilation present in continuous speech. The phonetic transcription includes word and syllable boundaries and explicit mark of changes in the transcription caused by coarticulation between words (reduction, elision). Phonemes are limited by spaces, syllables by '-' and words by '|'. If reduction between words makes the word boundary to be unspecified, a '/ /' symbol will be used to mark word limit (i.e. in Spanish "le escribo" pronounced without reduction " - l e – | - e s – " k r i – B o – " or pronounced with reduction between words " - l e/ /e s – " k r i – B o – " . The 'pause' between words has to be included in the phoneme 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 and is indicated by <pau>. If in the orthographic transcription a word is tagged as 'problematic' then it needs not to be transcribed phonetically (however the word boundaries have to be marked and the symbol * is included instead of the phonetic transcription). For expressive speech, filled pauses are transcribed in closest phonemes. Each producing partner has to provide the used phoneme set. If the corpus contains foreign sounds which cannot be represented by the language phoneme set then additional symbols have to be included.
Status	Mandatory except for Mandarin
Recommendation	Provide the guidelines for phonetic transcriptions in the LSP.
Comment	

Validation: The phonetic transcriptions of a sample of files are checked. A maximum of 5% errors is allowed. Details will be given in the separate validation manual.

2.7.1.3 Prosodic Transcription

Feature	Symbolic prosody annotation
Value	For baseline voices and cross-language conversion voices: - Phrase breaks are annotated using two levels: minor break (intermediate intonational phrase), major break (full intonational phrase). - Pitch accent (intonational prominence) is annotated using two levels: 'normal' or 'emphatic'. For tonal languages the value of the tones is annotated. The information about symbolic prosody complements the orthographic annotation: starting with the orthographic text, the symbol '#' is added after the words (without space between) with pitch accent (## for emphatic words). The breaks are included between words as <b> (minor break) or <BB> major break. Example: the #printer <b> run #out of #paper .<BB> The prosodic transcription should also be based on the acoustic signal.
Status	Mandatory
Recommendation	Provide the guidelines for prosodic transcription in the LSP.
Comment	None

Validation: A sample of files will be checked. A max. of 20% deviation on prosodic marks is allowed. Details will be given in the separate validation manual.

2.7.2 Segmentation

Feature	Phonetic segmentation
Value	For baseline voices and conversion voices: All baseline voices and conversion voices are segmented either automatically and/or manually. X% (20% for baseline; 5% for conversion; 0% on expressive speech.) of speech is checked manually for each sub corpus (C1.1,…, C3.3). The segmentation must match the manual phonetic transcription (if provided). Mandarin is segmented on syllable level. For each phoneme, 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. All the events (start and end positions) are indicated in seconds. The segmentation must match the manual phonetic transcription. For each phoneme, 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. 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 problematic cases 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 speech segments which have being labeled as “problematic” in a given word with a (*) symbol, do not need to be supervised.

Validation: The segmentations of speech segments are checked, equally distributed between the manual and automatic segmentation. A max. of 5% wrong segmentations is allowed for the manual part and 10% for the automatic part. An error in segmentation is defined as in terms of deviations in ms. Details will be given in the separate validation manual.

Feature	Word segmentation
Value	For expressive speech: All the expressive speeches are segmented either automatically and/or manually into words. For each word, the starting and ending time must be provided. If reduction of parts of words is produced, a ‘middle’ time between words is provided as starting or ending point. All the events (start and end positions) are indicated in seconds.
Status	Mandatory
Recommendation	None
Comment	No validation for expressive speech.

2.7.3 Pitch Marking

Feature	Pitch marks
Value	Many systems use pitch synchronous processing. For this reason speech signals of the baseline and conversion voices are labeled with pitch marks. Pitch marking point: consistent; points are defined with reference to the maximum of signal (maximum is defined in close neighborhood of the positive slope of laryngograph signal) A part (20% for baseline; 5% for conversion; 0% on expressive speech) of speech is checked manually for each sub corpus (C1.1,…, C3.3). Many systems use pitch synchronous processing. For this reason the baseline and conversion voices are labeled with pitch marks. The pitch marks are located at the instant of glottal closure as observed in the laryngograph channel. The pitch marks are determined for all the baseline voices and conversion voices. For the part of the baseline voices with supervised segmentation at least two hours are manually supervised.
Status	Mandatory
Recommendation	Although the use of a laryngograph makes 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 problematic cases 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 phonemes which have been labeled as “problematic” in a given word with a (*) symbol do not need to be supervised.

Validation: Maximum deviation from reference pitch mark: 5% of pitch period but not bigger than 0.5 ms, 3% of voiced/unvoiced errors; 3% voiced/unvoiced errors (automatic pitch ragger.)

2.8 Database interchange format

This chapter defines the database interchange format for the TC-STAR TTS speech databases.

2.8.1 Storage Media and Character set

Speech data will be stored on DVD's. An extra CD with only non-speech files (documentation, annotation, meta-files) will be considered. The character set should be UTF-8.

2.8.2 File Types

A complete database consists of signal and descriptive files. Signal files store the audio signals. 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 2.8.3), and a disk ID-file (DISK.ID, see 2.8.3).

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 label file will provide the annotation and transcription information.

Platform

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 96 kHz, 24 bit (16 bit optionally).

2.8.3 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 2.8.2, <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 2.8.1 – Content of root directory

README.TXT and COPYRIGH.TXT are formatted in the UTF 8. 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.

2.8.4 Speech and label file system hierarchy

The general structure for all signal and the label files is

```
/<database>/<type>/<speaker>/<subcode>
```

with "/" a generic file system separator symbol.

<database>	Defined as <dbName><#><language code> where <dbName> is TCTTS, <#> is 6 for TC-STAR, <language code> is <RFC 3066 code> RFC is the language code used as standard in HTML/XML files. 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 optionally the third 3-8 letters subtag in lowercase (if necessary to also disambiguate dialects within languages, e.g. RFC 3066 zh-CN for Mandarin, zh-CN-yue for Cantonese)[23]
<type>	Defined as <TTT> Where <TTT> is the type of the subcorpus, i.e. either BLR (Baseline), VCR (Voice Conversion), EXR (Expressive Speech)
<speaker>	Defined as <SCD> Where <SCD> is a two digits speaker ID as defined in section 2.8.9
<subcode>	Defined as <nm> Where n is the scenario code and m the subcorpus code, i.e. either 11 or 12 (for transcribed speech); 20 (for written text); 31,or 32, or 33 for selected phrases and 40 for expressive speech.

Table 2.8.2 – Directory structure

Validation: Implementation of correct file name system is checked.

2.8.5 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>/PROMPT	Prompt text
/<database>/SOURCE	Source code

Table 2.8.3 – Documentation file hierarchy

[23] This proposal follows the specifications of the LILA project (cf. http://www.lilaproject.org/)

Validation: Implementation of correct documentation file structure is checked.

2.8.6 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><speaker><subcode><NNNN>.<LL><F>
```

Where:

<dbID>	Database Identification Code (00-ZZ), for TC-STAR TTS: T6
<T>	Type, where B is baseline, V is voice conversion, E expressive speech here
<speaker>	Two digits speaker ID (SCD)
<subcode>	Two digits corpus subcode: 11, 12, 20, 31,32, 33, 40
<NNNN>	Utterance identification number: 0000-9999
<LL>	Two letters language code ISO 630-2
<F>	File type code: S: SAM label file L: pitchmark file from laryngograph recordings (text) P: phonetic segmentation file (text) W: words segmentation file (text) (for expressive speech) 1,2,3: speech signal files for channels 1 – 3

Table 2.8.4 – 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.

2.8.7 Speech file format

All signals are sampled synchronously with a sampling rate of 96 kHz, 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.

2.8.8 SAM Labels

Administrative data, speaker data, recording conditions and annotation data follows SAM label structure. This section describes the SAM labels used in the database. 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 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."
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.

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>/<speaker>/<sub code>	%s
SRC	Speech file names	A comma separated list of 11.3 file names, one per each speech recorded channel plus the laryngograph signal	%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: 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,CHN1 = %f, CHN3 = %f The SNR values (dBA) of the microphone recordings	%f
BWI	Bandwidth 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> with <gain>, <min>,<max> optional signal values; if they are not known, the values may be left empty, but the correct number of commas must remain. <prompt text> is the text that appears on the screen.	%d, %d, %d, %d, %d, %s
LBO	Orthographic transcription. One per each sentence in the Prompt text	BEG, MID, END,<sentence>	%d, %d, %d, %s
LBP	Prosodic transcription. One per each sentence in the Prompt text	BEG, MID , END,<sentence>	%d, %d, %d, %s
LBB	Phonetic transcription. One per each sentence in the Prompt text	BEG, MID, END,<sentence>	%d, %d, %d, %s
TXF	Name of the prompt text file	8.3 file name	%s
SYS	Labeling system	free form text	%s
SPA	SAMPA version	free form text	%s
ASS	Assessment code	free form text	%s

Table 2.8.5 TC-STAR TTS SAM labels

In Table 2.8.5 alternatives are enclosed in square brackets “[]” and separated by the vertical bar “|”.

Is possible for LBR, LBO, and LBP to include multiple sentences in one label file. Sentences can be delimited by using multiple LBR, LBO, LBP labels. Punctuation marks are also allowed within these labels.

2.8.9 SAM Label Files

In the following paragraphs the different types of label files are described.

Speaker File

The speaker file is:

spk<speaker>.txt where <speaker> is a two digits speaker ID as defined in Table 2.8.4

For each speaker in the database, a single file contains the basic information of the speaker. The file contains:

SAM Label	Description
LHD	Label header
SCD	Speaker code
SEX	Speaker gender
SNM	Speaker name (optional, only if the speaker agrees)
AGE	Speaker age
ACC	Speaker accent
PRF	Speaker profession
NL1	First native language
NL2	Second native language
ELF	End of label file

Table 2.8.6 Contents of the speaker label file

Recording platform file platf.txt

The file platf.txt describes the characteristics of the recording platform and recording environment. In principal, each database is recorded in a single place and a single equipment and the recording platform file is unique. If recordings are carried out in different places or platforms, there should be a file per recording site and should be properly documented.

SAM Label	Description
LHD	Label header
SAM	Sampling frequency
SNB	Number of (8-bit) bytes per sample
SBF	Sample byte order
SSB	Number of significant bits per sample
QNT	Quantization
NCH	Number of channels
REV	Reverberation
BWI	Bandwidth of speech signal
ELF	End of label file

Table 2.8.7 Contents of the recording platform label file

8.9.3 Annotation Files

There is one SAM label file assigned to each utterance (i.e. one for all the recording channels). Each SAM label file includes information about the recorded signal and its annotation.
The annotation labels are:

SAM Label	Description
LHD	Label header
CMT	Comment
DBN	Database name
DIR	Speech file directory
SRC	Speech file names
SCC	Scenario code
CCD	Corpus code
RED	Recording date
RET	Recording time
BEG	Labeled sequence begin position
END	Labeled sequence end position
NCH	Number of channels
SNQ	Signal/Noise Quality in dBA
LGG	Usefulness of Laryngograph signal
LBD	Label file body
LBR	Prompt text
LBO	Orthographic transcription
LBP	Prosodic transcription
LBB	Phonetic transcription
TXF	Name of the prompt sheet text file (optional)
SYS	Labeling system (optional)
SPA	SAMPA version (optional)
ASS	Assessment code (optional)
ELF	End of label file

Table 2.8.8 Contents of the annotation label files

Validation: 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.

2.8.10 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.
- <LL>W: contains the word transcriptions with time stamped segment boundaries.

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 phoneme 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 SAMPA symbol, 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.

The word segment boundaries are coded with one line for each word (including the symbol "<pau>" for pause). Each line consists on three fields separated by a tab stop. First, the word, 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 words from this file must match the words in the LBO field in the SAM label file.

Note: the precision of the time stamps should be enough to do not loose resolution when writing the file.

Validation: formal correctness of other label files is checked.

2.8.11 Table files

The table files are mandatory database files providing an overview of the TC-STAR 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

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 scenario (baseline, voice conversion) so that it allows to record the same speaker in more than one scenario, but with the same SCD.
SPEAKER.TBL contains the following fields:

SCD	Unique speaker code
SNM	Speaker name (if agreed by the speaker)
SEX	Speaker gender
AGE	Speaker age
ACC	Speaker accent
PRF	Speaker profession
NL1	First native language
NL2	Second native language

Table 2.8.9 – 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 (optional)

Table 2.8.10 – REC_COND.TBL mandatory fields

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

Validation: formal correctness of table files is checked.

2.8.12 Lexicon files

Lexicon Files LEXIC<nn>.XML

The lexicon table should be provided in the XML-format defined by LC-STAR [MOR 04]. An XML-based mark-up language was chosen to represent the linguistic information in a formal, unambiguous manner and easy to read. Moreover, the information can be processed by as many parties as possible. The XML parser that will be used to parse the lexica can be any XML version 1.0 compliant parser. Parsers should be able to deal with UTF-16.

As is defined in LC-STAR [MAL 04], lexica consist of a set of entry group elements.

- An entry group refers to a generic entry in a vocabulary. For each entry group, it is mandatory to specify:
  - orthography;
  - zero or more alternative spelling elements;
  - one or more entry or compound entry or abbreviation elements.
- An entry refers to one specific grammatical/morphological meaning of a vocabulary entry. For each entry, it is mandatory to specify:
  - One POS, together with its attributes. In case of multiple POS, or in case of multiple attributes of the same POS, multiple entries have to be specified within the same entry group.A description of all mandatory features and values for a given language will be provided in Design.doc
  - One lemma. It contains string data representing the entry lemma. In case of multiple lemmas, multiple entries have to be specified within the same entry group.
  - One phonetic transcription. It contains string data representing the entry phonetic transcription and syllabification. In case of multiple phonetic transcriptions, multiple entries have to be specified within the same entry group.
  - For application words, one APP tag has to be specified. The structure of APP tag is as follows:

    Subdomain_type1 No_of_entry 1

    …

    Subdomain_typeN No_of_entryN
- Compound entries will have the following structure:
  - phonetic transcription;
  - two or more entry elements (a subset of an entry), which are links to other entries. Each entry element must be characterized by an orthography and must contain one POS TAG, together with all of its attributes.
- Abbreviations from application wordlists will be tagged using the ABB tag. Each abbreviation must contain one or more EXP TAG and each EXP TAG contains
  - a string data representing the actual expansion (optional);

- one entry or compound entry element (mandatory).

- Each attribute has the default value NS (=Not Specified), which is always optional in the DTD (IMPLIED)..NS should be used only if the attribute is not mandatory in a given language.

In each entry the possibility of inserting a comment is also provided by the XML formalism <*!—insert here your comment* --> that can be used in any part of the Lexica.

No.	Spelling	POS	Lemma	Phonetisation + syllabification
1	**capitano (It)**	**NOM. Class:** *common.* **Number:** *singular*. **Gender:** *masculine.*	capitano	k a – p i - “t a – n o
		VER. Number: *plural.* **Person:** *3.* **Tense:** *present.* **Mood:** *indicative.* **Voice:** *active.*	capitare	“k a – p i – t a – n o
2	**dai (It)**	**VER. Number:** *singular*. **Person:** *2.* **Tense:** *present.* **Mood:** *imperative.* **Voice:** *active.*	dare	“d a – r e
3	**الحمراء (Ar)**	**NOM. Class:** *common.* **Number:** *singular*. **Gender:** *Feminine.*	عارمح	not available
		ADJ. Number: *singular*. **Gender**: *feminine.* **Case**: *Genitive.* **Degree**: *positive.*	عارمح	not available

Table 2.8.11. Logical structure for some entries in a Lexicon.

```
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE LEXICA SYSTEM "NewLexica7.dtd" >
<LEXICA xml:lang="IT">
      <ENTRYGROUP orthography="capitano">
            <ENTRY>
                  <NOM class="common" gender="masculine"
                        number="singular" />
                  <LEMMA>capitano</LEMMA>
                  <PHONETIC>k a - p i - " t a - n o</PHONETIC>
            </ENTRY>
            <ENTRY>
                  <VER tense="present" number="plural" person="3"
                        mood="indicative" voice="active" />
                  <LEMMA>capitare</LEMMA>
                  <PHONETIC>" k a - p i - t a - n o</PHONETIC>
            </ENTRY>
      </ENTRYGROUP>
      <ENTRYGROUP orthography="الحمراء" xml:lang="AR">
            <ENTRY>
                  <NOM class="common" gender="feminine"
                        number="singular" />
                  <LEMMA>عارمح</LEMMA>
                  <PHONETIC>not available</PHONETIC>
            </ENTRY>
            <ENTRY>
                  <ADJ case="genitive" degree="positive"
                        gender="feminine" number="singular" />
                  <LEMMA>عارمح</LEMMA>
                  <PHONETIC>not available</PHONETIC>
            </ENTRY>
      </ENTRYGROUP>
</LEXICA>
```

Table 2.8.12 XML-based coding of examples listed in Table 2.8.11

Packaging: Lexicon has to be split into two parts: proper names and common nouns. These parts should be split further into a set of smaller and more manageable files. Splitting is language dependent and must be done on an alphabetic base.

Filenames are

LEXIC <nn>.XML
where <nn> are two digit from 00 to 99.

Spliting criteria, filenames and mapping between filenames and content should be documented.

DTD File: LEXICON.DTD

A formally specified grammar (Document Type Definition or DTD), containing all the linguistic information described so far allows to validate automatically the XML-based lexica. The LEXICON.DTD file contains the DTD implementing the linguistic information described in LC-STAR [LC-STAR D2.x ref] All lexicons generated in LC-STAR uses a common DTD and is available at the web page of the project. For validation purposes, each partner should provide in the documentation (DESIGN.DOC) a Language Specific DTD to allow validation of mandatory attributes.
It should be noted [LC-STAR D2.x ref] that:

- Each lexicon and entry group has an optional attribute specifying the language and the attribute values of their sub-elements: we chose the standard XML attribute xml:lang whose possible values are defined by [IETF RFC 1766], *Tags for the Identification of Languages*, or its successor on the IETF Standards Track.

- For sake of simplicity, we associate each entry to as many triples (POS, Lemma, Phonetic Transcription) it can belong to, thereby allowing for repetitions in a triple.

- The characters range supported by the XML Standard [http://www.w3.org/TR/2000/REC-xml-20001006] is any UNICODE character, excluding the surrogate blocks, FFFE, and FFFF. Moreover, all the XML processors must accept UTF-16 encoding.

- However, special characters must be escaped by pre-defined strings when they are contained either in elements of type PCDATA (i.e. Parsed Character Data) or in attribute values of type CDATA (i.e. Character Data). The following Table illustrates the situation:

Special Character	Escaping String	Must be escaped in PCDATA Elements	Must be escaped in CDATA Attributes
&	&amp;	Yes	Yes
<	<	Yes	Yes
>	>	Yes (1)	Yes (1)
]]>	NOT AVAILABLE	No	Yes
‘	'	No	No
“	"	No	Yes

Table 2.8.12. Special characters that have to be escaped in PCDATA elements content and/or in CDATA attribute values.

Notes:

(1) According to the XML Standard “the right angle bracket (>) may be represented using the string "`>`", and must, for compatibility, be escaped using "`>`" or a character reference when it appears in the string "`]]>`" in content, when that string is not marking the end of a CDATA section.”

Validation: formal correctness of formats and DTD is checked

2.8.13 Documentation files

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

Root directory

The root directory contains three mandatory files:

COPYRIGH.TXT : a copyright text in ASCII format, mentioning also the TC-STAR 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.

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 2.8.13 – Allowed file formats in 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 information:
  - speaker selection procedure
  - accent regions of each speaker
  - a reasoned description of the regional pronunciation variants that are distinguished
  - age and sex of each speaker
  - native language, second native language and parents native language (if relevant)
- 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 (Unicode- UTF8)
  - any other language-dependent information such as abbreviations, proper name conventions, contractions July or July, isn't, cannot or can not, etc.)
  - markers for mispronunciations, recording truncations, unintelligible speech, non-speech acoustic events (*)and other language-specific symbols
- 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
  - analysis of frequency of occurrence of the speech segments (phonemes, diphones and triphones, syllables) represented in each subcorpora.
- Lexicon information:
  - Introduction
    - File Nomenclature and formats
    - Contents of the lexicon
    - Types of entries
    - Information contained in the lexicon (fields)
  - Format of the lexicon
    - Description of XML format used: (Entry_group, Entry, Entry_el…)
    - Description of fields
    - Splitting criteria
    - Description of DTD file. Mention LEXICON.DTD file
  - Morphological and syntactic information

  - POS set used, plus attributes and values mention LC-STAR deliverable D2.x
  - Description of multiple tagging possibilities
- Language specific guidelines
  - Principles of POS tagging and their attributes
  - Any systematic underspecification of POS attributes
  - Morphological boundaries
  - Principles of multiple tags
  - Choice of lemma
  - Instructions to the coder
  - Language Specific DTD
- Orthographical information
  - Character set used
  - Orthographical conventions used
  - Language specific orthography guidelines
  - Treatment of multi-token entries
  - Treatment of numeric entries
  - Treatment of acronym spellings
  - Treatment of abbreviations
  - Treatment of multiple spellings
- Phonemic transcriptions
  - Procedures used to obtain phonemic forms from orthographic input
  - SAMPA set used and that blanks separate individual phoneme symbols in the transcriptions
  - Use of syllable markers, stress markers (and tone markers or morphological markers, if provided)
  - Assimilation processes accounted for in the transcriptions
  - Treatment of multiple pronunciations
  - Treatment of foreign words: phonetization, tagging…
  - Language specific phonetic transcription guidelines to be included if relevant for the language
  - Stress markers and syllable boundary markers
  - Principles of assigning stress
  - Treatment of stress in multi-word entries
  - List of usually unstressed words
  - Tone markers
  - Morphological markers
  - Extra SAMPA symbols to cope with foreign languages
  - Mapping foreign phonemes to language specific phonemes
  - Phonetic transcription in names (nativization of foreign names, etc)

  - Dialect considered as ‘canonic’
  - Syllabisation conventions (i.e. geminates, etc)
  - list of PinYin syllables
- 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 labeling.

Character Table

It is mandatory that a sample character table in UTF 8 coding 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 2.8.12 gives an overview of documentation files:

Directory	File	Status
TABLE	REC_COND.TBL	mandatory
	SPEAKER.TBL	mandatory
DOC	DESIGN.DOC	mandatory
	UTF 8.PS	mandatory
	PLATFORM.DOC	optional
	SPELLALT.DOC	optional
	TRANSCRP.DOC	optional
	VALREP.{TXT|DOC}	mandatory
	LC-STAR. Deliverable D2.x	Mandatory

	PLATF.txt	Mandatory
	SPK<xx>	Mandatory
LEXICON	LEXIC<nn>.XML	Mandatory
	LEXICON.DTD	Mandatory
CORPUS	C11_T	Mandatory
	C12_T	Mandatory
	C20_T	Mandatory
	C31_T	Mandatory
	C32_T	Mandatory
	C33_T	Mandatory
	C4_T	Mandatory

Table 2.8.12 – Summary of Documentation files

2.8.14 Recommendations

Storage:

It is advisable to make backup copies to a safe media (e.g. DVD) as early as possible.

SAM label files:

It is forbidden to split items over more than one line to facilitate processing.

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

2.9 References

[ELB 04] Ellbogen,T; Schiel, F., Steffen,A (2004):'The BITS Synthesis Corpus for German', Proc. LREC2004

[MOR 04] Moreno, A. (2004) Specifications of lexicon interchange format. LC-STAR Technical report D3.0.

[KAI 01] Kain,A.; Macon, M.W.: 'Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction' Proc. ICASSP 2001

[BLA] Black, A.; Lenzo, K.: Building Synthetic Voices, FestVox 2.0 Edition ; http://www.festvox.org/

[MAL 04] Maltese, G. Montecchio, C. et al. General and language specific specification of contents of lexica. LC-STAR Project Deliverables D2. 2004

Appendices A and B

A1 Algorithms to Achieve High Triphone and Phoneme Coverage

According to Section 2.4.3, a high coverage of speech segments i.e. triphones or syllables within the entire speech corpora C_T is requested. Section A1.1 describes algorithms to achieve high coverage in triphones.
Furthermore, for the subcorpus C3.3_T, high frequencies of rare phonemes are requested. Each phoneme must occur at least 10 times. In section A1.2, the procedures to achieve this goal are described.

A1.1 Algorithm to Achieve High Triphone Coverage

The entire corpus of a particular language must contain more than 95% of the triphones of the language's LC-Star phonetic lexicon. For achieving this triphone coverage within C_T, the following databases are used:

- the corpora C1.1_T, C3.1_T and C3.3_T that are generated independently of triphone coverage aspects. Nevertheless, their contribution to the triphone coverage is to be taken into account. Therefore, the creation of these three corpora has to be finished before starting the generation of C1.2_T, C2_T and C3.2_T, in the following referred to as 'predefined corpora',
- the LC-Star phonetic lexicon of the considered language, in the following referred to as 'lexicon',
- a collection of short sentences from novels and short stories of the considered language consisting of more than one million running words, in the following referred to as 'novels' and
- the TC-Star EPPS corpus (Spanish / UK-English) or a similar text database for Mandarin, in the following, referred to as 'EPPS'.

In the beginning all available text material, (i.e., the predefined corpora, the novels, and the EPPS) is transcribed by a simple lexicon lookup; out-of-vocabulary words (OOV) s or words with ambiguous transcriptions are omitted. Now, we determine the baseline triphone coverage achieved through the predefined corpora by dividing the set of different triphones in the predefined corpora (T_{pre}) by that in the lexicon (T_{lex}).
First the corpus C2_T is generated. The ext corpus is based on novels and short stories and it is considerably smaller than the corpus C1.2_T which is based on transcribed parliamentary speeches. (EPPS). We apply a greedy algorithm selecting in each iteration that sentence from novels for which the transcription maximizes the ratio between the number of triphones not yet covered by the already extracted sentences in the considered sentence and the number of words thereof. i.e., in the first iteration, we go through all considered sentences s_n (n=1,...,N) and compute the ratio r_n between the number of triphones in s_n not yet covered by T_{pre} and the number of words in s_n. Finally, we select that sentence $\tilde{n}_1$ that maximizes the ratio r_n. $\tilde{n}_1$ is removed from the set of considered sentences and its triphone contribution is added to T_{pre}. The algorithm iterates until the number of extracted words is 27000 or more (cf. Section 2.1.4).
Second, the text corpus C1.2_T is generated based on EPPS. The procedure is the same as for the creation of C2_T, cf. above. The algorithm iterates until the number of extracted words is 36000 or more (cf. Section 2.1.4).
At last we have to check whether the already achieved triphone coverage satisfies the coverage criterion (at least 95%). If not, we apply the same greedy algorithm as described above taking the lexicon itself as input material and interpreting each entry as a sentence s_n. The greedy algorithm is iterated until the triphone coverage criterion is satisfied. Now, sentences are searched in several databases (for instance in the Internet) that contain these selected lexicon entries. The collection of

the found sentences is a starting point for the generation of the corpus C3.2_T. The remaining part of C3.2_T can be used for optimizing other coverage criteria, cf. Section 2.4.3.

A2 Mimic Sentences Adaptation and Diphone Sentences (C_10SR)

A2.1 Mimic Sentences: Calibration of the Template Speech

According to Section 2.3.3.4 for intra-lingual voice conversion speakers of the same language (the corpus speakers) try to mimic the timing and the accentuation pattern as well as the pitch contour of the given template utterances. It is important that all selected corpus speakers mimic the same template speaker as the voice conversion parameter training is optimized in this way. Since in general the average fundamental frequency of the corpus speakers can differ essentially from that of the template speaker (e.g., due to different genders, age, etc.), the template utterances have to be adapted in such a way that the corpus speakers feel comfortable mimicking the given pitch contour. Therefore the template utterances are manipulated by means of a PSOLA technique that changes the fundamental frequency by adding a positive or negative speaker-dependent offset while keeping the speaking rate and the voice characteristics.

A2.2 Generation of the Diphone Sentences (C_10SR) from the corpus C_200SR

According to section 2.4.6.2 of the specification, 10 sentences need to be synthesized and to be additionally recorded (C_10SR, as a natural reference). The synthesis is performed using the annotated corpus C_200SR.

Out of the recorded and annotated 200 sentence corpus, 10 new sentences for testing the acceptability of potential concatenative synthesis using this voice are constructed. The test scenario is presuming an example of a diphone approach and TD-PSOLA-based concatenation.

Each sentence or phrase should have a maximal word overlap of 70% to each of the 200 sentences and should contain at least 5 syllables.

Text analysis and construction:

For this purpose all words of the corpus C_200SR are clustered into sub corpora (according to their main word classes given by POS information). The sub corpora are sorted with regard to the word frequency. In parallel, all diphones of the corpus C_200SR are analyzed and sorted in the order of their frequency.

Basing on this limited diphone list and on the sub corpora of the words, an expert is manually constructing (recombining) new phoneme sequences (by connecting diphones which do not come from the same phonemic context). The main target is to generate as many new words as possible by using the most frequent diphones and by avoiding frequent words from the sub corpora. Considering the validation target, in average, each third ‘new word’ should be generated from completely independent phonemic contexts, not seen in the original corpus.

Using the new and partly-new phonemic words, the expert is manually constructing 10 sentences with the mentioned restrictions and basing on the standard grammar of the language. Preferably, sentences or phrases of mean length (with regard to the original corpus) should be generated. Nonsense utterances should be avoided if possible.

Corpus and synthesis preparations:

The 10 sentence text corpus is given by the orthographic form of the corresponding phonemic word sequences. It is recorded with regard to the standard conditions used for the corpus C_200SR and provides a natural reference. Furthermore, diphone synthesis by using the mini diphone inventory (analyzed and compiled from the corpus C_200SR) and a standard TD-PSOLA algorithm is processed.

Remarks:

By theoretical consideration, the achievable synthesis quality is strongly limited, since the corpus speaker is not forced to speak in a monotone manner. Therefore, many concatenation units do not match and alternative units are not available within this mini inventory. Nevertheless, the resulting

synthesis chunks can provide a first impression of the acceptability of speech synthesis basing on the selected voice.

B1 Noise, Frequency Range, Reverberation and Recording

B1.1 Frequency Range

The range of the speech signal should be 40 Hz – 20 kHz. Signals with frequencies outside of this range are caused by non speech. No post filtering is accepted to deliver the database

B1.2 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)

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).
Back ground 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 2.1.1. This definition includes the hearing characteristic of the human ear.

B1.2.1 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[24]. Based on these values SNR_A is defined as:
$SNR_A = dBA_S - dBA_N$

[24] A Matlab program calculating dBA by approximating the filter of Fig B1 is available at Siemens. This program has been made available by the acoustic group (Prof. Hugo Fastl) of the Institute of Man-Machine Communication at the Technical University of Munich (TUM).

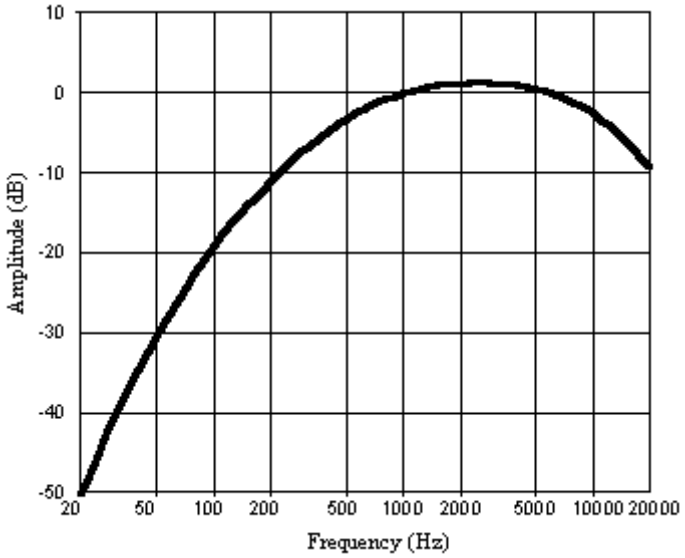

Fig B1.2.1 - Transfer function of the filter for measuring dBA.

B1.2.2 NC-xx[25]

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 determinine 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).

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

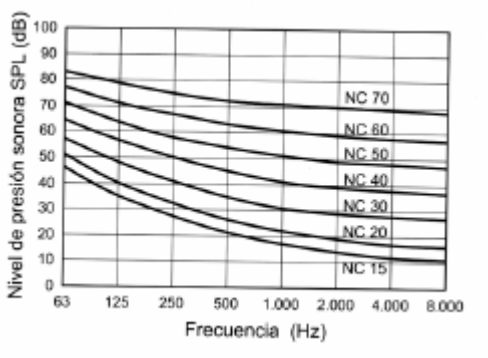

Figure B1.2.2 - NC curves

B2 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 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).

Validation: A max. of 5% of the label files may have label REV > 0.3
A max. of 2% of the label files may have label REV > 0.4

B3 Recording

In the following proposals for recording hardware and software are given however each partner is free to use whatever best fits and is in accordance with the specifications.

B3.1 Proposals for recording software

- SpeechRecorder
- M-audio FireWire (http://www.m-audio.de/ or www.m-audio.com) + driver from ASIO
- Samplitude 7.0 or higher (http://www.samplitude.com)
- Steinberg Cubase or Nuendo
- Apple LOGIC PRO 7
- NannyRecord

B3.2 Proposals for recording hardware

- RME HDSP 9652 Audio Card, ca. 499 EUR, (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)
- RME Fireface 800 sound card

B3.3 Proposals for large membrane condenser microphone

- Neumann Type TLM 103 (600 – 1000€)
- Microtech Gefell M-930 (founded by Georg Neumann) (600 – 1000€)
- Microphone "Rode K2", ca. 999USD (http://www.fullcompass.com/Products/pages/SKU--65126/index.htm)

B3.4 Proposals for the laryngograph

- EG2-PC (2 channel electroglottograph with 35 mm electrodes; (www.glottal.com)
- Laryngograph processor (www.laryngograph.com).

B3.5 Proposals for the close-talk microphone

- Cardioid condenser microphone (suppresses most surrounding noises)
- Sennheiser ME104
- Omni condenser microphone
- ShureWBH53B.

3 Specifications of Evaluation of Speech Synthesis

3.1 Introduction

The development of speech technology in TC-STAR is evaluation driven. Assessment of speech synthesis is needed to determine how well a system or technique compares to others or how it compares with previous version of the system. In order to make a useful diagnose of the system in TC-STAR we will not only make a test of the whole component but also specific tests for each module of the speech synthesis system. In this way we can assess better the progress on specific modules. Furthermore, it allows identifying the best techniques in the different processes that are involved in speech synthesis. To allow the comparison of different modules we have defined a common specification of the modules and specific test.

In order to increase the critical mass in speech synthesis and the impact of the evaluation, the TC-STAR project opens the evaluation on speech synthesis to external partners. The needed language resources produced at TC-STAR are shared with partners willing to participate in the evaluation. With this goal, the TC-STAR partners involved on Work package 3 have founded ECESS, the European Centre of Excellence on Speech Synthesis [ECESS]. This Centre is open to any partner willing to participate on the evaluation and to share the technology for research purposes. This document includes the result of discussion with ECESS partners.

Although some objective metrics have been proposed to evaluate some modules of the speech synthesis system, in most of the cases the evaluation relies on human judges. The evaluation can be carried out using a web interface. This makes possible that the subjects can perform the test from their home computer that has to be equipped with a high-speed internet connection, a standard sound card and closed headphones. For each language, between 15 and 20 subjects participate in the evaluation. At every evaluation campaign, the sampling rate of the synthetic speech will be defined (recommended: 24 kHz).

The outline of this document is as follows. Section 3.2 of this document describes the three modules of the speech synthesis systems, functionalities and an overview of the interfaces. Section 2.3, defines the evaluation tests for modules. Section 3.4 focuses on the evaluation of specific research activities: voice conversion and expressive speech. Section 3.5 defines the evaluation of the speech synthesis component (whole system).

3.2 Definition of speech synthesis modules

Text-to-speech systems perform a range of processes, from text normalization, pronunciation, several aspects on symbolic and acoustic prosody, etc. Finally we are interested on the *quality* of the overall system. However, the evaluation of the whole (*black box evaluation*) does not allow pinpointing which part of the system causes the most relevant problem. Furthermore, this method does not allow participating on the evaluation to small teams of researchers whose specialty of research is in one specific topic. In TC-STAR we will certainly evaluate whole systems (see section 3.5), but we also want to evaluate different tasks to drive more valid conclusions about the results of different algorithms. Defining *modules,* with well defined input and output allows keeping constant all the modules except one and comparing the results caused by the algorithms involved on that module (*glass box* evaluation).

There are many processes involved in speech synthesis. Researchers working in a particular one would prefer to make specific tests to evaluate that process. For instance, some tests have being proposed to evaluate each aspect of prosody, from intonation, pausing, accentuation, etc. However, from a pragmatic point of view, when designing a general evaluation framework, the number of modules needs to be limited. The evaluation of speech synthesis involves in many cases human evaluation and is needed to limit the number of test for each campaign. Also, in order to compare different systems only generic modules can be defined because not all the systems are built up of

the same processes. Furthermore, although we assume that speech synthesis is built up of independent modules, in fact this is not absolutely true. For instance, a promising area of research is modeling the correlation between the different features related with prosody (f0, duration, etc.). Keeping these processes together allows modeling this interaction.
Therefore, there is a compromise in the number of modules. In TC-STAR we define three broad modules: symbolic preprocessing, prosody generation and acoustic synthesis. The modules have been defined through their interfaces, i.e., the formal description of the input and output [ecess_iface].

Symbolic preprocessing.
The first module has to perform three tasks:

- Word normalization: the input text is transformed into words that would be found in common lexica with complete coverage. For instance, dates, postal address, numbers, abbreviations, are transformed into these words. One of the tasks of this module is the tokenization. This includes, for languages like Mandarin, word segmentation. It also includes correct tokenization of punctuation and detection of the end of the sentence.
- Pronunciation: the pronunciation of each word is derived. The representation depends on the language and also on the used technology. For English and Spanish and most of western languages, the pronunciation is represented using the phonetic transcription, including lexical stress and syllabic boundaries. For Mandarin, the pronunciation is represented using syllables (therefore, syllables boundaries are not needed), including the lexical tones.
- POS tagging: each word is tagged with the disambiguated POS (Part-of-Speech).

The text input to the TTS system will be formatted into a SSML [SSML] conforming document. SSML (Speech Synthesis Markup Language) is a W3C Recommendation, a XML-based markup language specifically designed for assisting the generation of synthetic speech. SSML defines tags to control the text structure (paragraphs and sentences) and to give information about the desired prosody and style (voice gender, speaker age, specific processor voice name, emphasis, pitch contour, duration, etc.). Since all the tags defined in SSML are optional, plain text can easily be transformed into a SSML document by embracing it between `<speak>` and `</speak>` SSML definition tags. In fact, during evaluation, no control marks will be included in the text.
In the second and third evaluation campaign, the mark-up language will be extended to fit the research results of the project. It is expected that prosody and segmental information will be derived from the source voice and included in the input text using appropriated mark-up. This information will be used during in speech synthesis.
Basically, the output of the module is a sequence of words. For each word, information about POS and *pronunciation* is included. The words present in the output will be those originally present in the input text and those resulting from the normalization of dates, abbreviations, acronyms, etc.
As a general rule, the POS are coded using the tagset defined in LC-STAR (only the main tag, not the attributes). However, for each language, a different tagset can be defined depending on the available resources. In particular, it has been agreed that in the first evaluation campaign, the English tagset will be the defined in the Penn Treebank [Marcus93].
For English and Spanish, the phonetic transcription will be coded using the SAMPA phonetic alphabet [SAMPA]. The phonetic transcription includes syllable boundaries and lexical stress. The phonetic transcription corresponds to words uttered in isolation. Determining the pitch accents of

the sentences or applying coarticulation rules is postponed till module 2. For Mandarin, the pronunciation is represented using the syllables including the tones.

Prosody generation

This module generates the acoustic prosody representation for the sentence. In the first phase of the project, the acoustic prosody is specified by means of F0-countour, intensity contour and energy contour. Optionally, the interface supports other parameters related to voice quality or symbolic prosody. These optional parameters will not be used in the evaluation of module 2 during the first phase of the project.

The prosody generation module associates to each word a list of corresponding phonemes. This may not necessarily be equal to the phonetic transcription itself, since assimilation of vowels, creation of diphthongs and similar phenomena may be considered. For Mandarin, the state of the art systems are based on syllable so each word is represented by a list of syllables.

For each phoneme (syllables for Mandarin) the following information is mandatory:

- The duration of the phonemes (syllables in Mandarin) is expressed in milliseconds
- The fundamental frequency of the phonemes (syllables for Mandarin) is expressed as pairs of time_position (in milliseconds) and f0_value (in Hz). This is a flexible method allowing for several approaches for frequency specification to be used in the system. The simplest case is one single value for the whole phoneme, allowing, for instance, the common approach of indicating the frequency value at the middle of the phoneme. More detailed fundamental frequency curves can be implemented by sampling this curve with different resolutions according to the prosody model. For unvoiced phonemes this value is irrelevant.
- The power contour or intensity is specified in a similar way, using pair's time_position and power_value. The energy is expressed by the mean power in dB. For a given pair (time_position, power_value), the value measures the power from the last time_position to the actual time_position. For instance, to give the mean power for each phoneme, the ending position of the phoneme should be used as time_position.

The description of prosody using only duration, f0 and intensity is not complete. For instance, it is known that the prominence of the syllable is correlated with the spectral distribution of the energy. Therefore, many systems use symbolic information to select units that are more appropriated. Therefore, the interface definition allows specifying some phonological information as the stress and the presence of intonation break. However, in this version of the evaluation specifications, in the sake of simplicity, these parameters are ignored.

As stated above, the prosody information is expressed adding information to the phoneme. The information associated to the syllable is added to the first phoneme of the syllable. We have preferred not to impose a hierarchical representation (word composed of syllables and syllables composed of phonemes) to allow the representation of syllables that span from the end of a word to the beginning of next word. Adding syllable information to the first phoneme of the syllable preserves the information about syllable boundaries.

The accent level will be labeled with positive integers indicating the importance of the accent (1 indicates normal, 2 indicates emphatic). If the syllable is the last of the word, information about the break index tier will be added. The break level is specified using the categories defined in the standard markup language SSML [2]: none, x-weak, weak, medium, strong and x-strong.

As a summary, for each word, the phonemes (syllables for Mandarin) of the word are related. Then for each unit, the following information is added:

- Information about segmental duration, f0 contour and intensity contour
- For each syllable include symbolic information. For English/Spanish, specify information about the pitch accent (normal or emphatic). The information is added to the first phoneme of the syllable. For Mandarin, the Pinyin transcription system already includes tone information.

- For the last syllable of the word, information about the break index tier is included. If the units are phonemes (English/Spanish), this information is included in the first phoneme of the syllable..

Acoustic synthesis
The last module produces synthetic speech based on the prosody representation. For evaluation a *MS-WAV* file will be produced. The baseline systems are based on concatenative speech using unit selection from a speech database but the defined interface allows other synthesis methods.

3.3 Evaluation of the speech synthesis modules

3.3.1 Module 1: Text analysis

The goal of the text analysis is to transform the orthographic input string to the representation of the sounds. It involves text normalization, which transform ambiguous text such as numbers, dots and abbreviations into non-ambiguous words (which are known as "standard words"). In the case of Mandarin, this module segments the character stream into words. This module also copes with grapheme to phoneme conversion and with the assignment of lexical stress and syllable boundaries. Furthermore, this module tags the words with the POS (part-of-speech label), which is needed for prosody assignment.
Therefore, the input of the module is orthographic text and the output consists on three layers: standard words, their phonetic representation and the POS-tags. In TC-STAR we adopt the criteria defined in LC-STAR [LC-STAR D2]. Given the orthographic text, the output of the module is not unique because sometimes the same text can be read using different words (normalization ambiguity) and also because some words accept several pronunciations. The module has to produce one particular output and the evaluation metrics has to accept it, if it is one of the correct ones.
The correct pronunciation of words depends on prosody. Specifically, depending on the pausing and on the speaking rate assimilation to adjacent words can occur. However, usually the evaluation of grapheme to phoneme it is based on isolation words. One of the reasons is that the rules for word pronunciation in context are believed to be straightforward [EAGLES, pp. 514]. For sake of simplicity, in TC-STAR the pronunciation algorithm will be evaluated on isolation.
The input and output format of the following test is the one defined for the text analysis module. Therefore, in any test, the output includes the orthographic transliteration, POS tagging and phonetic transcription. The evaluation agency should select the needed information for each specific test.

Test M1.1: Text Normalization.

1. The evaluation agency selects N_1 running words presented in paragraphs or sentences. The words are selected from:
   - 50% from the domain C3.2 *(frequent phrases),* in Section 2.
   - 25% from text transcriptions from the parliamentary domain[26].
   - 25% from text transcriptions from the parliamentary domain, formatted according to the rules of the translation engine (WP1)
2. Each system processes the text and produces a tokenized version of the text using only "standard words".
3. The evaluation agency produces the reference transcription following the TC-STAR convention. If more than one transcription is acceptable then they have to be coded in a format compatible with the evaluation tool (either lists of transcriptions or graphs coding

[26] In this document, *parliamentary domain* means transcriptions from the European Parliament. In case that this is not applicable for a given language (e.g. Mandarin), a similar domain will be defined.

the acceptable transcriptions). In some cases, the normalization requires deep knowledge of the domain, as in the case of expand technical abbreviations. In this case, one possible expansion is the one that will be used for a native speaker without specific knowledge of the domain.

4. For each system, the output of module 1 is evaluated, taking into account insertions, deletions and substitutions. The figure used to assess the systems is the word error rate, as defined in the tool provide by NIST to asses continuous speech recognition.

Test M1.2: Word Segmentation (Mandarin)

1. The evaluation agency selects N_1 running words presented in paragraphs or sentences. The words are selected from the domain defined in LSP for Mandarin.
2. Each system processes the text and produces a tokenized version of the text including word boundaries.
3. The evaluation agency checks the word segmentation. The figures used are precision and recall.

Test M1.3: Evaluation of POS-tagger.

1. The evaluation agency select $>N_2$ running words. Approximately 50% is from the parliament domain and 50% of other general domains (for instance, news). The evaluation corpus is split into paragraphs. The text should be selected so that it is not expected to have problems in the tokenization.
2. Each system tags the text.
3. The evaluation agency creates the reference tagged text using the POS-tags defined in LC-STAR (only the main tag, not the morphologic attributes). For UK-English, as the LC-STAR lexicon is not available, the tagset is the defined in the Penntree-bank project will be used.
4. For each system, the output of module 1 is evaluated comparing the output with the reference. The metric is the percentage of errors. In the case that the number of tags is different to the reference, the evaluation agency will report this and will supervise the alignment (for instance, deleting the problematic sentences) to be able to count the errors.

Test M1.4: Evaluation of grapheme-to-phoneme

1. The evaluation agency selects $> N_3$ words. Half of the words are derived from the parliamentary domain and 50% from other domains. The words are classified in:
   - Common words
   - Geographic places such as towns, countries, etc.
   - Name of persons (family and given names) and organizations.
2. Each system produces the pronunciation for the selected words, the lexical stress, and the syllable boundaries.
3. The evaluation agency produces the reference pronunciation, including pronunciation alternatives if needed.
4. For each system, the evaluation agency computes the pronunciation word error rate: in case there is an error in one or more phonemes in the word, the pronunciation is not correct. Analogously, the evaluation agency computes the error rate, at the word level, for lexical stress and syllable boundaries. The evaluation should be given for the different domains

and inform about the percentage of unknown words (with respect to the lexicon used in TC-STAR).

5. The foreign names will not be taken into account when computing the error.[27]

3.3.2 *Module 2: Prosody.*

The output of the second module is *acoustic prosody.* Many systems predict *symbolic prosody* as a first step to produce the acoustic parameters. However, while many research laboratories use the same values as acoustic prosody (pauses, f0 contour, segmental duration, energy contour), the coding of symbolic prosody in some cases depends on the theory behind the models. Furthermore, some experiments reveal that the evaluation at the symbolic level cannot substitute the acoustic tests [1, pp. 518].

Usually, acoustic objective measures are used to evaluate models and to estimate their parameters. For instance, to have a first evaluation of the segmental duration model usually the MSE (mean square error) is used. This metrics compares, for each phoneme, the prediction of the model with the duration measured in a human sentence (reference speech). However, the correlation between these objective measures and the perceptual judgment is not very high (for instance, in melody modelling, a straight line can give acceptable results in terms of correlation and MSE but the synthetic speech is monotonous). Therefore, in order to evaluate prosody we rely on judgment tests of prosody.

To assess this module all the systems under comparison will share both the input and the backend. The input will consist on normalized words and the correct pronunciation, stress, POS and syllable boundaries, as detailed in the interface definition. For the output, the same backend will be applied. The backend will consist in re-synthesis of natural sentences. This avoids distortions that can occur in synthetic speech. Therefore, the assessment of naturalness and quality of intonation is easier.

To do the speech re-synthesis a toolkit as Praat [Praat] will be used to change the prosody (f0, duration and energy) according to the output of the module under evaluation. This requires that the natural utterances are segmented into phonemes. The toolkit will be agreed before each campaign by the partners. All the partners will be able to use it during development.

The natural sentences should be uttered by the baseline speaker. If this is not possible (speakers are not available), other speaker should be selected taking into account that the pitch mean is the same that the one form the speaker producing the baseline voices. The selection should also consider voice quality after F0 manipulation using pitch synchronously labeled speech units. The speaker has to be instructed to speak at the same mean speaking rate that the one in the baseline voice.

The evaluation of the prosody will be based on paragraphs. The evaluation of sentences in isolation usually gives better but unrealistic results with respect to their use in continuous speech as is the case of broadcast news and parliamentary speeches (applications in TC-STAR).

To evaluate the prosody we will focus on “naturalness of the prosody: intonation, rhythm, etc.)”. The subjects are instructed not to take into account noises or acoustic distortions. The systems are evaluated using an absolute scale going from 1 (very unnatural) to 5 (completely natural).

One major problem in the evaluation of prosody is the influence of the segmental component on speech perception. In TC-STAR we will adopt the BLURR method proposed by Sonntag and Portele [Son98]. This method generates delexicalized utterances were the lexical information is lost and only the melody and temporal structure is presented. The basic idea is to generated, for voiced sounds, a harmonic signal using only the first and second harmonics. The amplitude of the second harmonic component is one fourth with respect the first one. The signal is generated taken into account the prosodic description (f0, duration and energy). The unvoiced sounds are reflected like pauses. Based on these delexicalized utterances, two tests are proposed. A judgment test is used to

[27] The pronunciation of foreign proper names is out of the scope of the project and will not be considered in the first phase of TC-STAR. It is expected that frequent foreign names will appear in the lexicon (LC-STAR proper names or baselines voices). However, the main goal of the test M1.3 is not to evaluate lexicons but methods to cope with out-of-vocabulary words.

rate if the prosody is appropriated to a given test. Furthermore, a functional test is defined so that the subjects have to choose the text sentence more appropriated to a given delexicalized utterance.
For judgments tests, the number of subjects and items to be rated depends on the number of systems. The general recommendations are:

- All the systems produce synthetic speech based on the same items (paragraphs).
- At least S=20 subjects participate in the evaluation. Each subject evaluates all the systems unless the number of systems to be evaluated is too big (>20).
- Each subject listens to each paragraph only once.
- Each system should receive at least 40 ratings.
- For each subject, the presentation order is randomized both in systems and in items to eliminate any dependency on the order.

Table 3.1 specifies the number of subjects and number of items as a function of the number of systems. The last column shows the number of ratings to be done for each subject.

Y = #Systems	*#Subjects*	*#items*	*#ratings/subject*	*#ratings/system*
1	20	6	5	100
2-6	20	Y*3	Y*3	60
7-10	20	Y*2	Y*2	40
>10	2*Y	20	20	40

Table 3.1: Number of subjects and items as a function of the number of systems to be evaluated

Test M2.1: Evaluation of prosody (using segmental information).

1. The evaluation agency selects N_4 items (paragraphs) from the parliamentary domain, distributed over different *melodic domains* taking into account the real distribution: declaratives, questions, list, etc. The number of items depends on the number of systems (see Table 3.1).
2. For each item, the evaluation agency produces the input to module 2 (words, POS, pronunciation, syllable boundaries, lexical stress).
3. Each system produces the prosody description for these items. In the first campaign, one additional baseline system will be produced as reference.
4. The evaluation agency generates synthetic speech based on the prosody description using the generation toolkit.
5. A judgment test is performed by naive subjects. Each subject has to rate in a scale of 5, the naturalness of the voice. The subjects are instructed to pay attention to prosody (not speech quality or noise). To avoid the learning effect, each subject judges each item produced only by one system. The number of subjects depends on the number of systems to be evaluated (see Table 3.1).

Test M2.2: Judgment test using delexicalized utterances.

1. The evaluation agency produces the input to module 2 in the same way that defined in test M2.1 (cf. Test M2.1, points 1 and 2).
2. Each system produces the prosody description for these items.
3. The evaluation agency generates delexicalized utterances based on the prosody description: voiced sounds, the signal is generated using two harmonic sinusoidal functions; unvoiced

sounds are rendered as silence. The f0 and energy (for voiced sounds) and duration (for voiced and unvoiced) are consistent with the prosody description.

4. A judgment test is performed by naive subjects. Each subject reads the original text sentences and judge is the prosody is good or not for that sentence, using a 5 points scale.

Test M2.3: Functional test using delexicalized utterances.

1. The procedure for generating the stimulus (delexicalized utterances) is the same than in test M2.2.
2. For each utterance, the subjects have to choose which sentence, from a set of 5, is more appropriated to that prosody. The sentences should differ either in phrase modality, boundaries, phrase accent or number of syllables. One of the sentences is the correct one, i.e., the original sentence presented to each system.

3.3.3 Module 3: Speech generation.

The third module produces speech from the phonetic and (acoustic) prosody description. Segmental quality or segmental identification is one of the main factors in getting good overall quality, in particular in words which cannot be easily predicted from the context as is the case of proper names, figures, etc.

Some decades ago several tests were designed to evaluate segmental quality. For instance, Diagnostic Rhyme Test (DRT) and the Modified Rhyme Test (MRT), evaluate the segmental intelligibility by identification of the initial or final consonants of CVC words from a close set of options (two or six). However, these tests are not very suited to evaluate state-of-the-art methods. Corpus based systems look for speech segments as longer as possible with prosodic restrictions. To synthesize short words prosody is not so relevant and therefore the system could be tuned to a very different working mode. The system could find the CVC words found in the database. Therefore, the quality would be near the same than natural speech. The *SAM Standard Segmental Test* is other method that evaluates the segmental quality using meaningfulness words. As stated before, the unit selection systems are designed for reading sentences, no words, and the segmental quality depends on the task. Therefore, we propose to evaluate segmental quality based on sentences.

One important effect of segment quality is intelligibility. Furthermore, other aspects, as naturalness is also affected by segment quality. Intelligibility can be evaluated by functional tests (subject transcript what they listen) but naturalness requires judgment tests similar to the ones needed to evaluate the overall quality of the system. Both aspects, intelligibility and naturalness, need to be evaluated because sometimes there is a trade off between them. It is possible to produce speech perfectly intelligibly but very unnatural (for instance choosing a very slow speech rate). So, both tests (functional and judgment) will be used to evaluate segment quality.

Intelligibility: functional test.

To evaluate the intelligibility we will use the Semantically Unpredictable Sentences test (SUS), based on the one proposed by SAM. This test consists of a set of syntactic structures or templates. The lexical slots are filled with words from the parliament domain. The words should be chosen so that they are known for most of the people (for instance not technical names too specific). The sentences are semantically unpredictable but syntactically correct. For each sentence, the input to the speech generation module is prepared. Based on that, the systems produce synthetic voice which is presented to the subjects. They transcribe what they listen. The measure is the word error rate.

Judgment

Segmentation quality will also be evaluated based on judgment. Some sentences will be selected from the parliamentary domain. For each sentence, the input to the speech generation module will

be produced. Based on this input, all the systems produce the synthetic speech. The synthetic sentences are presented to the subjects who rate the sentences in terms of naturalness and intelligibility.
Note that in both test, the correct prosody has to be provided. The most problematic part is the prosody. The prosody should be as natural as possible. To do that, the prosody features (acoustic prosody) will be based on the reading of the sentences by a professional speaker. However, the acoustic modules are tuned to one particular speaker. This is especially true for concatenative systems based on unit selection. In order to provide to the modules with a prosody description matched with their voices, if it is possible, the sentences will be read by *baseline speakers*, i.e., the speakers that same speakers that recorded the baseline corpus. If this is not possible, a speaker with the same tone will be used. In this case, the evaluation agency will provide with an adaptation corpus.

Test M3.1: Evaluation of speech generation module: functional test

1. The evaluation agency selects N sentences to be used as templates. These sentences should be syntactically correct. The length of the templates is approx. 10-15 words.
2. For each template, several sentences are produced changing the lexical words (names, adjectives, verbs, etc.) by other words, with the same morphosyntactic features chosen from the parliament domain.
3. For each sentence, the evaluation agency produces the input to the module 3: words, phonetic transcription, and prosody description. This is based on the reading of the sentences by one professional speaker which should be the same that the baseline voice speaker. If this is not possible (speaker is not available) then a speaker with the same tone and speech rate will be selected and an adaptation corpus will be provided.
4. Each system generates the synthetic speech based on the input. The systems have to produce the words in the input but they are not forced to respect exactly the input features (phonetic or prosodic). Some systems use the prosody description only as a rough guide but the final prosody depends on the selected segments.
5. The synthetic utterances are degraded with an additive or multiplicative noise so that the intelligibility of the synthetic speech drops. The noise characteristics will be included in the evaluation report.
6. The evaluation agency presents the synthetic voice to the subjects. The subjects transcribe in words the listened sentence. The sentences can be listened twice. The evaluation metrics is the word error rate as used in speech recognition (nist tool).

Test M3.2: Evaluation of speech generation module: judgment test

1. The evaluation agency selects N items (sentences) from the parliamentary domain. The length of the templates is approx. 10 words.
2. For each sentence, the evaluation agency produces the input to the module 3: words, phonetic transcription, and prosody description. This is based on the reading of the sentences by one professional speaker which should be the same that the baseline voice speaker. If this is not possible (speaker is not available) then a speaker with the same tone and speech rate will be selected and an adaptation corpus will be provided.
3. Each system generates the synthetic speech based on the input. The systems have to produce the words in the input but they are not forced to respect exactly the input features (phonetic or prosodic). Some systems use the prosody description only as a rough guide but the final prosody depends on the selected segments.
4. The evaluation agency presents the stimuli to the subjects. The stimuli are the synthetic sentences. Furthermore, if the *baseline speaker* was available, the recordings are added as a top-line reference.

5. The subjects rate the naturalness and the intelligibility of the sentence in a scale from 1 to 5.

3.4 Evaluation of specific research topics

The evaluation described previously refers to conventional text-to-speech systems. It is needed significant improvement to achieve natural speech. Furthermore, in TC-STAR it is planned to investigate on two specific research areas: voice conversion and expressive speech. These two topics require specific evaluation test.

3.4.1 Voice conversion (VC)

Voice conversion is the adaptation of the characteristics of a source speaker's voice to those of a target speaker. When evaluating voice conversion technology, generally, we have two questions in mind:

- Does the technique change the speaker identity in the intended way?
- How is the overall sound quality of the converted speech?

The answers can be found applying subjective and objective error criteria. The former is based on listening tests. The latter expresses the distance between the converted speech and corresponding reference speech of the target speaker. However, our experience shows that the objective evaluation of voice conversion technology features severe shortcomings. Consequently, in this document, we develop a plan limited to subjective measures.
In TC-STAR, both conventional intralingua and cross-language voice conversion are to be investigated. The considered languages are English, Spanish and Mandarin, the combinations for cross-language voice conversion are English-Spanish and English-Mandarin.

The Training Corpus
As stated in the specifications on LR, the voice conversion corpus consist of four bilingual speakers (two female and two male). Each speaker produces about one hour of speech of both covered languages. The read contents are based on parallel texts taken from parliamentary speeches.

The Evaluation Corpora
For subjective evaluation, we found that none of the conventional procedures provides the information required for completely answering the first above question. Therefore, we suggest an evaluation method to be used in TC-STAR that, in some respects, is based on a proposal of Kain and Macon [Kain01]. Having a look at state-of-the-art voice conversion technology, we note that most of the systems only transform vocal tract and excitation whereas some approaches aim at transforming the speaker-dependent prosody as well. To be applicable to both kinds of systems, we propose to create two separate evaluation corpora that exclude or include prosody conversion, respectively.

The Evaluation Corpus Excluding Prosody. In order to achieve a similar prosody of all involved speakers, we apply an extension of the 'mimic' approach presented in [Kain01]. In the first evaluation the voice conversion is applied only to the speakers specified in this deliverable (D8a). In the next evaluation campaigns, new speakers will be recorded only for the evaluation.
The mimic voice conversion corpus contains 200 sentences. These sentences are split into two sets, development corpus (150 sentences) and evaluation corpus (50 sentences).
In the first evaluation campaign 4 voice transformations are defined:

- One transformation from a female voice into a female voice.
- One transformation from a female voice into a male voice.
- One transformation from a male voice into a female voice.

- One transformation from a male voice into a male voice.

For the evaluation of English intralingual conversion, we choose those four speakers that have the most native-like pronunciation. For cross-language conversion to English, we take those speakers that have the source language as mother tongue.
In second and third evaluation campaign, new speakers will be developed for evaluation. The development corpus has to be provided to estimate the transformation and the evaluation corpus will be used for testing the performance (evaluation tests).

The Evaluation Corpus Including Prosody. Here, we expect the corpus speakers to use their individual prosody, i.e., no template speaker is required.

Subjective Evaluation
In order to prevent the subjects from interpreting their decisions, they should not be familiar to the background of the test. In particular, they must not know the contents of this evaluation plan. I.e., ideal evaluation subjects are persons that do not have specific knowledge about speech processing at all.
The evaluation web page contains a clear instruction of what the subjects are to do, e.g.:

> *We are analyzing differences of voices. For this reason, you are asked to identify if two samples come from the same person or not. Please, do not pay attention to the recording conditions or quality of each sample, only to the identity of the person.*
> *So, for each pair of voices, do you think they are*
> *(1) definitely different,*
> *(2) probably different,*
> *(3) not sure,*
> *(4) probably identical,*
> *(5) definitely identical? ”*

Voice Identity Conversion
To keep the evaluation task as convenient and clear as possible, two speech samples are presented at a time. Each speech sample consists of 10 sentences that are randomly chosen from the evaluation corpus consisting of 50 sentences. The subjects are not forced to listen to the complete sample but can stop the playback whenever they want. The samples of two compared voices are based on identical sentences, whereas, for each comparison, the randomization is executed anew to prevent the subjects from becoming bored. Each subject evaluates the same test, i.e., the randomizations are executed beforehand. The evaluated voice conversion system has to convert the determined 10 sentences using the four defined transformations. During the evaluation, the subjects listen to 4 voice pairs consisting of the conversion results and the respective reference (target) speech. Besides, they have to rate the similarity of the unconverted voices, i.e., we have 4 more pairs that consist of the source speech and the reference. These 8 voice pairs are randomized, thus the subject does not know if he compares the converted voice with the source or the target.

Preparing the Test. During the recording of the evaluation corpus excluding prosody, we adjust the pitch of the template speaker by adding a pitch offset in the way that the respective corpus speaker feels comfortable. To make the prosody as speaker-independent as possible, in the test, this offset is to be deducted. This is done by providing the evaluated voice conversion system with the values of the pitch offset of the source and target speaker for each considered pair of utterances in the test. As each voice conversion system should include a pitch modification facility, this pitch offset is to be taken into account when synthesizing the converted speech. When comparing unconverted source and target utterances, the mean pitch of the source speech is adapted to that of the target speech by means of a PSOLA technique. A deterioration of the speech quality could be accepted as the subjects are asked to ignore it when evaluating the voice identity conversion.

Voice Conversion Score. In order to compare the performance of different voice conversion systems or to control a system's progress from one evaluation to the next, we define a voice conversion score that has to have similar properties as the mean opinion score used for quality assessment. Since the performance of the conversion highly depends on the difference of the involved voices (source and target), this score should take into account both the distance between the converted and the target voice and that between source and target voice. Here we define a score to measure the *subjective distance* between the target speaker and the transformed speaker, which takes values between 0.0 and 1.0.

- Let's be s(converted, target) and s(source, target) the rate of the subject when comparing the converted and target voices and source and the target voices. The value of s(.) goes from 1 (voice conversion success) to 5 (voice conversion failure).
- For each subject, and for each transformation, we define:

  D_s = [5-*s*(converted,target)]/[5-*s*(source,target)]

Note that

- If *s*(converted,target) < *s*(source,target), one should set D_s=1.0 per definition.
- If *s*(converted,target) = *s*(source,target)=5, the sample should not be counted.
- In the other cases, this equation becomes 1.0 if *s*(converted,target) = *s*(source,target), i.e., if the conversion showed no progress.

The final voice conversion score is the mean over all considered samples of all involved subjects.

Overall Speech Quality

Since it is widely used in telecommunications, for measuring the quality of the converted speech, we apply a mean opinion score test [ITU.P800]. The listeners are asked to assess certain sentences according to the following scale: (1) bad; (2) poor; (3) fair; (4) good; (5) excellent. The mean opinion score is the arithmetic mean of all subjects' individual scores.

Test Definition. To determine the best achievable conversion quality, the eight voices contained in the training and in the evaluation corpus are also considered. For the test, they are mixed up with the 16 conversion outputs

Test VC.1: Evaluation of research on voice conversion excluding prosody

1. The voice conversion will be evaluated in several languages (English, Mandarin and Spanish). For each language, 4 voice transformations are defined: female-to-female, male-to-male, male-to-female, female-to-male,. In the first evaluation campaign, the speakers are included in the TC-STAR voice conversion corpus. For the following campaign, two new speakers are added. (This can be either intra-lingual or cross-lingual, according to the evaluation schedule)
2. The new speakers will read 200 sentences selected from the voice conversion corpus. The sentences will be read using the *mimicking style*, as defined in the LR specifications. (D8a).
3. From these sentences, 150 will be used as training data and will be sent to the partners. The other 50 will be used for evaluation purposes.
4. Each voice conversion system transforms 50 sentences using the four defined transformation.
5. The subject rate if a given voice pair come or not from the same person. The number of pairs is 4 pairs to compare the target and the converted voices and 4 pairs to compare the target and the source voices. Obviously, the pitch of the transformed voice should be similar to the pitch of the target voice. Furthermore, the pitch of the original source voice is shifted to be similar to the pitch of the target voice

For each pair, the subjects listen two files containing 10 evaluation sentences (they are not required to listen all the sentences) and rate the identity of both voices from 1 (definitely identical) to 5 (definitely different).
The subjects are also asked to assess certain transformed sentences using a mean opinion score test.

Test VC.2: Evaluation of research on voice conversion including prosody

This test is very similar to the test VC.1. The only difference is that all the original sentences (source and target) are uttered using the natural prosody of the speaker. Furthermore, the source voice is presented without pitch adjustment.

3.4.2 Evaluation of research on expressive speech (ES)

Most of the evaluation procedures in expressive speech are functional tests related with emotion: synthetic speech is produced using one of a given predefined set of emotions. The subjects are asked to identify the emotion on the speech (close set answer). The aim of TC-STAR is not to produce emotional speech but expressive speech. One characteristic of expressive speech is that it can signal para-linguistic information using prosody (in the broad sense).
Produce expressive speech from general text requires very high knowledge of the world and high cognitive capabilities. However, in TC-STAR we want to explore how some para-linguistic information can be derived from the source speech and used to produce the synthetic voice.
It is difficult to establish a general functional test for expressiveness even for restricted spoken styles like the ones found in the parliament. In this first design we rely on a judgment test related with the expressiveness. The subjects will be asked about the degree of expressiveness. Furthermore they will judge if the speech is appropriated.
To evaluate the expressive speech component in the TC-STAR framework we require that the items include broad linguistic context. If a sentence or paragraph is presented without context it may be difficult to infer the attitude of the speaker and other affects. This general statement holds both for generating the synthetic speech (systems) and for evaluating the synthetic speech (judges).
Furthermore, one research direction requires the source speech in order to extract information from the source speech. This speech is transcribed and translated using the same conventions that in TC-STAR WP1 and WP2. For each word in the transcription of the source speech, the starting and ending time are provided. In this first stage of the research we propose to use recorded speech (not real speech from the parliament): a professional speaker reads a paragraph taken from the parliament. This allows to record a training/adaptation corpus and to get recording with the highest quality.

Test ES: Evaluation of research on expressive speech.

1. The evaluation agency prepares the inputs using the following procedure:
   - Selects N=8 documents (transcription of one complete interventions from a parliament) in the source language (English). Take also the original speech and the translation into Spanish.
   - For each document in the source language, select one paragraph.
   - Record the paragraph by one speaker which is able to imitate the original speech from the parliament. The reading should reflect what appear in the transcription, avoiding big speech disorders (repetitions, reformulations, etc.). The speaker can be either one speaker from the voice conversion speakers or other professional speaker. In this case, an adaptation corpus has to be provided.
   - Label the source speech using the WP1 orthographic conventions. Furthermore, state the starting and ending times for each word.

- The input to the speech synthesis system is a) the text of whole document in the target language (linguistic context); b) the text of the selected paragraph in the source and target language; c) the reading of the selected paragraph, in the source voice and the labeling.

2. Each system produces synthetic speech related to selected paragraph in the target language. As bottom-line, one of the systems is the baseline system (without introducing features for expressive speech).
3. M subjects (M=20) evaluate the synthetic speech from all the pairs paragraph-system. The subjects are presented a) the document in the target language and b) the synthetic speech from all the speakers. For each signal they have to answer the following questionnaire:

   Q1*: A given voice is expressive if it transmits not only the content but also feelings of the speaker, or position of the speaker with respect what is being said or about the listener, or which part is more relevant, etc. Please listen to the following speeches and judge the expressiveness of the voice:*

   a. The voice is not specially expressive
   b. The voice is slightly expressive but not appropriated in this context
   c. The voice is very expressive but not appropriated in this context
   d. The voice is slightly expressive and appropriated in this context
   e. The voice is very expressive and appropriated in this context

   Q2: *Rate from 1 to 5 the following statement: The prosody (intonation, speed, etc.) is natural and appropriated along the paragraph.* (1: absolutely disagree; 5: completely agree**).**

3.5 Evaluation of the speech synthesis component

In order to evaluate the system as a whole we will use a black box test. Subjects are asked to indicate their subjective impression of global quality aspects of synthetic output by means of rating scales. The evaluation protocol will be based on the ITU-P85 recommendation [ITU-P85]. In particular, we will follow the recommendations of a recent review of ITU-P80 [CS]. As in the other cases, the best system on each evaluation campaign will be used as baseline for the next campaigns..

In the first evaluation campaign, this test will be used to define the baselines systems. The test will not evaluate voice conversion and expressive speech. Therefore no information about the source speech or speaker will be provided. In the second campaign this test will be redefined taking into account the experience on the first campaign with tests VC (voice conversion) and ES (expressive speech).

Test S1: Evaluation of speech synthesis component

1. The evaluation agency selects N_4 items (paragraphs) from the parliamentary domain. The number of items depends on the number of systems (see Table 3.1). Half o the items will be represented using the normal orthographic convention found in the parliamentary transcriptions. The rest will be represented following the conventions of the output of the spoken translation module (WP1).
2. Each system produces synthetic voice.
3. A judgment test is performed by naive subjects. Each subject has to rate in a scale of 5, several aspects of the voice. To avoid the learning effect, each subject judges each item produced only by one system. The number of subjects depends on the number of systems to be evaluated (see Table 3.1).

3.6 Bibliography

[EAGLES] Handbook of Standards and Resources for Spoken Language Systems. Edited by Dafydd Gibbon, Roger Moore and Richard Winski; Walter de Gruyter Publishers, Berlin & New York, 1997

[ECESS] European Center of Excelence on Speech Synthesis, www.ecess.org.

[ITU.P85] ITU-T Recommendation P.85, "A method for subjective performance assessment of the quality of speech output devices", International Telecommunications Union publication 1994.

[ITU.P800] Methods for Subjective Determination of Transmission Quality," ITU, Geneva, Switzerland, Tech. Rep. ITU-T Recommendation P.800, 1996.

[Kain01] A. Kain and M. W. Macon, "Design and Evaluation of a Voice Conversion Algorithm Based on Spectral Envelope Mapping and Residual Prediction", in Proc: of ICASSP'01, Salt Lake City, USA, 2001.

[Marcus93] M. P. Marcus, B. Santorini, M. A. Marcinkiewicz: *Building a Large Annotated Corpus of English: The Penn Treebank*, in Computational Linguistics, Volume 19, Number 2 (June 1993), pp. 313--330 (Special Issue on Using Large Corpora).

[SAMPA] SAMPA computer readable phonetic alphabet, http://www.phon.ucl.ac.uk/home/sampa/home.htm

[Sonntag98] G. P. Sonntag, T. Portele, "*PURR - a method for prosody evaluation and investigation*", Journal of Computer Speech and Language, Vol.12, No.4, October 1998 Special Issue on Evaluation in Language and Speech Technology, 437-451

[SSML] D. C. Burnett, M. R. Walker, A.Hunt, "Speech synthesis markup language (SSML) version 1.0," W3C Recommendation, Sept. 2004. http://www.w3.org/TR/speech-synthesis/

4 XML Interface Specification

4.1 Introduction

One of the objectives of TC-STAR is to design a modular synthesis system consisting of three main modules: symbolic pre-processing, prosody generation and acoustic synthesis. The modules will be implemented by different partners, and existing components will require some adaptation to fit into this modular system. Thus, a common definition of the interfaces for inter-module communications is required.

In the following sections we provide a description of the system input requirements (Section 4.2), the interfaces between the text processing and the prosody generation modules (Section 4.3), and the prosody generation and acoustic synthesis modules (Section 4.4). All the information will be formally coded in XML using the DTD included in Section 4.6. Only one DTD will be used in our modular paradigm, each module filling the corresponding part of the XML document.

For the sake of simplicity and efficiency, we will take advantage of the formal definitions achieved in the LC-STAR [Mal 04] project. We will use DTD coded there to implement the POS tagging formal definitions. This DTD is included in Section 4.7, and it is also available at the homepage of the project [Mal 04]

This document includes several examples to clarify the different parts of the problem and to illustrate some practical considerations. Whenever possible, examples have been inserted in the text of the corresponding section (e.g. the SSML examples in section 4.2). However, long examples showing documents written in XML using the TC-STAR DTD have been included in the appendixes at the end of this report (Section 4.8). In particular, sec.4.8.1 shows an input SSML document using different available SSML tags, the corresponding TC-STAR XML document that will be feed into the prosody module appears in Section 4.8.2, and the corresponding input to the synthesis systems is shown in Section 4. 8.3.

```
<?xml version="1.0" encoding="UTF-16"?>

<!DOCTYPE tts PUBLIC "TC-STAR"
          "TC-STAR.dtd">
<tts xml:lang="es">
 <s>
  <TOKEN token="Cuerpo">
   <WORD word="Cuerpo">

    <POS>
     <NOM class="common" number="singular" gender="masculine"/>
    </POS>

    <PHONETIC>kw'eR - po</PHONETIC>

   </WORD>
  </TOKEN>

  <prosody rate="-20%">
   <TOKEN token="gaseiforme">
    <WORD word="gaseiforme">
     <POS>
      <ADJ number="singular" gender="neuter" type="qualitative"/>
```

```
    </POS>
    <PHONETIC>Ga-sej-f'or-me</PHONETIC>
   </WORD>
  </TOKEN>
 </prosody>

 <TOKEN token="que">
  <WORD word="que">
   <POS> ... </POS>
   <PHONETIC> ... </PHONETIC>
  </WORD>
 </TOKEN>

 <TOKEN token="sin embargo">
  <WORD word="sin">
   <POS> ... </POS>
   <PHONETIC> ... </PHONETIC>
  </WORD>
  <WORD word="embargo">
   <POS> ... </POS>
   <PHONETIC> ... </PHONETIC>
  </WORD>
 </TOKEN>

 <TOKEN token="ofrece">
  <WORD word="ofrece">
   <POS> ... </POS>
   <PHONETIC> ... </PHONETIC>
  </WORD>
 </TOKEN>

 <TOKEN token="resistencia.">
  <WORD word="resistencia">
   <POS> ... </POS>
   <PHONETIC> ... </PHONETIC>
  </WORD>
  <WORD word=".">
   <POS> ... </POS>
   <PHONETIC> ... </PHONETIC>
  </WORD>
 </TOKEN>

 </s>
</tts>
```

4.2 System input

The text input to the TTS system will be formatted into a SSML [Bur 04] conforming document. SSML (Speech Synthesis Markup Language) is a W3C Recommendation, a XML-based markup language specifically designed for assisting the generation of synthetic speech.

SSML defines tags to control the text structure (paragraphs and sentences) and to give information about the desired prosody and style (voice gender, speaker age, specific processor voice name, emphasis, pitch contour, duration, etc.). Since all the tags defined in SSML are optional, plain text can easily be transformed into a SSML document by embracing it between <speak> and </speak> SSML definition tags.

Here we present two illustrative examples showing some of the capabilities of the markup language, as they appear in [Bur 04].

4.2.1 SSML example 1

```
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE speak PUBLIC "-//W3C//DTD SYNTHESIS 1.0//EN"
                  "http://www.w3.org/TR/speech-synthesis/synthesis.dtd">
<speak version="1.0"
  xmlns="http://www.w3.org/2001/10/synthesis"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
  http://www.w3.org/TR/speech-synthesis/synthesis.xsd"
       xml:lang="en-US">
  <p>
    <s>You have 4 new messages.</s>
    <s>The first is from Stephanie Williams and arrived at <break/> 3:45pm.
    </s>
    <s>
      The subject is <prosody rate="-20%">ski trip</prosody>
    </s>

  </p>
</speak>
```

4.2.2 SSML example 2

```
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE speak PUBLIC "-//W3C//DTD SYNTHESIS 1.0//EN"
                  "http://www.w3.org/TR/speech-synthesis/synthesis.dtd">
<speak version="1.0"
  xmlns="http://www.w3.org/2001/10/synthesis"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
  http://www.w3.org/TR/speech-synthesis/synthesis.xsd"
  xml:lang="en-US">

  <p>
    <voice gender="male">
      <s>Today we preview the latest romantic music from Example.</s>

      <s>Hear what the Software Reviews said about Example's newest hit.</s>
    </voice>
  </p>

  <p>
    <voice gender="female">
      He sings about issues that touch us all.
    </voice>
  </p>

  <p>
    <voice gender="male">
      Here's a sample.  <audio src="http://www.example.com/music.wav"/>
      Would you like to buy it?
    </voice>
  </p>
```

```
</speak>
```

4.3 Interface: Text processing – Prosody generation

The Text Processing Module is in charge of the tokenization, POS tagging and phonetic transcription of the input text. A *token* is an individually distinguishable element of the input text. Each token is divided in *words*, each of them having an associated *transcription* and one *POS tag*. The words present in a document will be those originally present in the input text, those resulting from the normalization of dates, abbreviations, acronyms, etc., and those included by the different modules (text normalization or prosody generation) as a result of different necessities (e.g. *fillers* in case of expressive speech).

Phonetic transcription information will be coded for each word as the way the word is spoken in isolation. We will use the SAMPA phonetic alphabet with syllable boundary marker (-), stress marker (“) and tone markers for tonal languages. The codification of the POS is the one used in the LC-STAR project, and all the details can be found in [Mal 04]. For each POS different attributes can be marked (e.g. number, person, case, mood, ...) and the `not_specified` default attribute is always implied.

In section 4.5, we provide a hierarchical structure illustrating the XML tags (<TOKEN>, <WORD>, etc.) and its relation with each other. Refer to section 4.8 to see a working XML example using the TC-STAR DTD (note that the example is provided for illustration or the different tags only, it does not necessarily need to be correct).

4.4 Interface: Prosody generation – Acoustic synthesis

The prosody generation module will fill in the information regarding the syllabic structure of the words, corresponding phonemes, and their associated pauses, accents, fundamental frequency and voice-quality attributes. As before, section 4.5 clarifies the hierarchical structure of the XML tags (<PHON>, <SYL>, <voice-quality>, etc.) and its relation with each other clearly. Refer to section 4.8.3 to see a working XML example using the TC-STAR DTD (note that the example is provided for illustration or the different tags only, it does not necessarily need to be correct).

```
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE tts PUBLIC "TC-STAR"
            "TC-STAR.dtd">
<tts xml:lang="es">
 <s>
  <TOKEN token="Cuerpo">
   <WORD word="Cuerpo">

    <POS>
     <NOM class="common" number="singular" gender="masculine"/>
    </POS>

    <PHONETIC>kw'eR - po</PHONETIC>

   </WORD>
  </TOKEN>

  <prosody rate="-20%">
   <TOKEN token="gaseiforme">
```

```
    <WORD word="gaseiforme">
     <POS>
      <ADJ number="singular" gender="neuter" type="qualitative"/>
     </POS>
     <PHONETIC>Ga-sej-f'or-me</PHONETIC>
    </WORD>
   </TOKEN>
  </prosody>

  <TOKEN token="que">
   <WORD word="que">
    <POS> ... </POS>
    <PHONETIC> ... </PHONETIC>
   </WORD>
  </TOKEN>

  <TOKEN token="sin embargo">
   <WORD word="sin">
    <POS> ... </POS>
    <PHONETIC> ... </PHONETIC>
   </WORD>
   <WORD word="embargo">
    <POS> ... </POS>
    <PHONETIC> ... </PHONETIC>
   </WORD>
  </TOKEN>

  <TOKEN token="ofrece">
   <WORD word="ofrece">
    <POS> ... </POS>
    <PHONETIC> ... </PHONETIC>
   </WORD>
  </TOKEN>

  <TOKEN token="resistencia.">
   <WORD word="resistencia">
    <POS> ... </POS>
    <PHONETIC> ... </PHONETIC>
   </WORD>
   <WORD word=".">
    <POS> ... </POS>
    <PHONETIC> ... </PHONETIC>
   </WORD>
  </TOKEN>

 </s>
</tts>
```

4.4.1 Phonemic and syllabic information

The prosody generation module will associate to each word a list of corresponding phonemes. This may not necessarily be equal to the phonetic transcription itself, since assimilation of vowels, creation of diphthongs and similar phenomena will be considered.

We will add all the information related to a syllable to the first phoneme of that syllable. This methodology allows for the disassociation of words and syllables, and phonemes of different words can be easily associated to the same syllable (particularly useful in case of the association phenomena, for instance).

We will use information regarding the **beginning of a syllable** (the first phoneme of each syllable will have an extra XML element representing the syllable), and whether it is the **last syllable** of a word (in which case a flag will be present in the syllable structure to indicate it; absence of this flag means non-final syllable).

In order to label the break index tier, we will follow the guidelines set by SSML [Bur 04], where five categories are defined: `none`, `x-weak`, `weak`, `medium`, `strong` and `x-strong`.

The accent level will be labelled with positive integers indicating the importance of the accent (`1` indicates *primary* accent, `2` indicates *secondary*, and so on).

4.4.2 Intensity, duration and frequency

Each phoneme will have reference **duration**, as estimated by the prosody module, expressed in **milliseconds** and will apply to the whole phoneme. The synthesis module is required to match this duration as close as possible.

The **fundamental frequency** of the phonemes will be expressed as pairs of **time_position** (in `milliseconds`) and **f0_value** (in `Hz`). This is a flexible method allowing for several approaches for frequency specification to be used in the system. The simplest case would be one single value for the whole phoneme, allowing, for instance, the common approach of indicating the frequency value at the middle of the phoneme. More detailed fundamental frequency curves can be implemented by sampling this curve with different resolutions according to their needs. Each sampled point will be specified as the time value where it occurred, and the frequency value.

The **energy** or **intensity** of the phoneme will be specified in a similar way. As a measure of intensity we will use the mean power in `dB`, which is defined for a segment of the phoneme as: 10·log(1/N · sum(s[n]^2)) where *s*[*n*] is the digitized speech segment of length *N*. Each segment will be defined with a time position indicating where it ends. The start of the segment does not have to be explicitly specified, since it can be obtained from the end position of the previous segment. In case of specifying the energy of the whole phoneme, or dealing with the first segment, the start time position will be taken as zero.

4.4.3 Voice Quality

As there is no widely accepted definition of *voice quality*, we will contemplate different options to incorporate this knowledge into the synthesis system. The approaches proposed in subsections 4.4.3.1 and 4.4.3.2 closely follow the studies of Laver [Lav 80] as presented by E. Keller in [Kel --]. Subsection 4.4.3.3**¡Error! No se encuentra el origen de la referencia.** is based on the voice-source parameterization studies by C. Gobl [Gob 03]. Voice quality information should be considered **optional** since not all synthesis procedures require this knowledge.

4.4.3.1 Laryngeal parameters

In order to select different voice properties using articulatory correlates or the larynx and associated muscles, the following voice-forms are available:

- **modal**,
- **falsetto**,
- **whisper** (can be combined with falsetto or modal),
- **creak** (can be combined with falsetto or modal),
- **harshness** (can be combined with other voice-forms),

- **breathiness** (can be combined with other voice-forms).

In order to create new voice-forms, more than one element can be used, indicating its relevance with a percentage of the total (e.g. modal 50%, creak 15% and breathiness 35%).

4.4.3.2 Tension settings

Different qualities can also be indicated by using the following classification of voices based on a general tensing or laxing of the entire vocal tract musculature:

tense	sharp	shrill
metallic	strident	lax
soft	dull	guttural
	mellow	

Only one voice type can be selected at a time, they can not be combined as in section 4.4.3.1

4.4.3.3 Glottal source related parameters

In order to determine the voice quality of the synthetic output, we will use the following features related to voice-source [Gob 03]:

- **EE** excitation energy (overall strength of source excitation) [dB],
- **OQ** open quotient (proportion of the glottal period during which the glottis remains open) [%],
- **AS** aspiration noise [dB],
- **RA** return time (sharpness of glottal closure) [%],
- **RG** glottal frequency (degree of boosting in the areas of the first and the second harmonic) [%],
- **RK** glottal asymmetry (relation between opening and closing phase of the glottal period) [%].

EE and AS are power measures and will be expressed in dB (see 4.4.2). OQ, RA, RG and RK are expressed as fractions of the pitch period (in %).

4.5 Interface structure

Here it follows a structured description of the data involved in the TTS process, and its internal organization in the document. A formal DTD description is included in Section 4.6. Optional elements are explicitly marked below (no mark indicates the element is mandatory).

Please note that the *OPTIONAL* mark in `voice quality` is used to indicate that voice quality information is a completely optional information that does not need to be used at all in the TTS system. On the other hand, the *OPTIONAL* mark in `PHON` merely indicates that either the word results in no associated phonemes, or that the text has not yet passed through the prosodic module. `PHON` is the basic information unit of the synthesis module, necessary to produce any speech output.

1 TOKEN

a) ***WORD*** ("word")

- POS (see [1])
- PHONETIC ("transcription")
- PHON ("phoneme") **[FILLED BY TEXT-PROCESSING MODULE]**
  - Duration (milliseconds)

- Frequency (sampled curve)
  - Time pos (milliseconds), value (Hz)
  - …
- energy
  - time pos (milliseconds), value (Hz)
  - …
- voice-quality **[OPTIONAL]**
  - laryngeal **[OPTIONAL]**
    - ⇒ modal [%]
    - ⇒ falsetto [%]
    - ⇒ whisper [%]
    - ⇒ creak [%]
    - ⇒ harshness [%]
    - ⇒ breathiness [%]
  - tension **[OPTIONAL]**
    - ⇒ tense, sharp, shrill, metallic, strident, lax, soft, dull, guttural or mellow
  - source **[OPTIONAL]**
    - ⇒ EE [dB]
    - ⇒ OQ [%]
    - ⇒ AS [dB]
    - ⇒ RA [%]
    - ⇒ RG [%]
    - ⇒ RK [%]
- SYL **[ONLY IN FIRST PHONEME OF SYLLABLE]**
  - Last-syllable (boolean flag)
  - Accent level (1, 2, etc.)
  - Break level: duration (milliseconds) and strength (none, x-weak, weak, medium, strong or x-strong)

- PHON ("phoneme") **[FILLED BY TEXT-PROCESSING MODULE]**
- …
- PHON ("phoneme") **[FILLED BY TEXT-PROCESSING MODULE]**

b) WORD

c) …

d) WORD

2 TOKEN

3 …

4 TOKEN

4.6 TC-STAR DTD

```
<?xml version="1.0" encoding="UTF-16"?>
<!--
        TC-STAR TTS DTD
        Specification of the interfaces in the TTS modular system.

        Based on the SSML DTD as provided by W3C. The original copyright
        is reproduced below.
-->

<!--

SSML DTD (20031204)
```

```
Copyright 1998-2003 W3C (MIT, ERCIM, Keio), All Rights Reserved.

Permission to use, copy, modify and distribute the SSML DTD and
its accompanying documentation for any purpose and without fee is
hereby granted in perpetuity, provided that the above copyright
notice and this paragraph appear in all copies.

The copyright holders make no representation about the suitability
of the DTD for any purpose. It is provided "as is" without expressed
or implied warranty.

-->

<!-- External Entities (BEGIN) -->
<!ENTITY % lc-star   SYSTEM "NewLexica12.dtd">
%lc-star;
<!--  External Entities (END) -->

<!-- Entity Declaration (BEGIN) -->
<!--ENTITY % ns "not_specified"-->
<!ENTITY % tension      "tense | sharp | shrill | metallic | strident |
                         lax | soft | dull | guttural | mellow | %ns;">
<!ENTITY % boolean      "true | false | %ns;">
<!ENTITY % accent       "primary | secundary | %ns;">

<!-- SSML Entities (BEGIN) -->
<!ENTITY % duration "CDATA">
<!ENTITY % integer "CDATA">
<!ENTITY % uri "CDATA">
<!ENTITY % audio " TOKEN | audio ">
<!ENTITY % structure " p | s">
<!ENTITY % sentence-elements " break | emphasis | mark | phoneme | prosody | say-
as | voice | sub ">
<!ENTITY % allowed-within-sentence " %audio; | %sentence-elements; ">
<!-- SSML Entities (END) -->

<!-- Entity Declaration (END) -->


<!-- Root element -->
<!ELEMENT tts (%allowed-within-sentence; | %structure; | lexicon | metadata |
meta)*>
<!ATTLIST tts
       xml:lang NMTOKEN #REQUIRED
>

<!ELEMENT p (%allowed-within-sentence; | s)*>
<!ATTLIST p
       xml:lang NMTOKEN #IMPLIED
>

<!ELEMENT s (%allowed-within-sentence;)*>
<!ATTLIST s
       xml:lang NMTOKEN #IMPLIED
>

<!ELEMENT voice (%allowed-within-sentence; | %structure;)*>
<!ATTLIST voice
       xml:lang NMTOKEN #IMPLIED
       gender (male | female | neutral) #IMPLIED
       age %integer; #IMPLIED
       variant %integer; #IMPLIED
       name CDATA #IMPLIED
>
```

```
<!ELEMENT prosody (%allowed-within-sentence; | %structure;)*>
<!ATTLIST prosody
       pitch CDATA #IMPLIED
       contour CDATA #IMPLIED
       range CDATA #IMPLIED
       rate CDATA #IMPLIED
       duration %duration; #IMPLIED
       volume CDATA #IMPLIED
>
<!ELEMENT audio (%allowed-within-sentence; | %structure; | desc)*>
<!ATTLIST audio
       src %uri; #REQUIRED
>

<!ELEMENT desc (#PCDATA)>
<!ATTLIST desc
       xml:lang NMTOKEN #IMPLIED
>

<!ELEMENT emphasis (%allowed-within-sentence;)*>
<!ATTLIST emphasis
       level (strong | moderate | none | reduced) "moderate"
>

<!ELEMENT say-as (#PCDATA)>
<!ATTLIST say-as
       interpret-as NMTOKEN #REQUIRED
       format NMTOKEN #IMPLIED
       detail NMTOKEN #IMPLIED
>
<!ELEMENT sub (#PCDATA)>
<!ATTLIST sub
       alias CDATA #REQUIRED
>
<!ELEMENT phoneme (#PCDATA)>
<!ATTLIST phoneme
       ph CDATA #REQUIRED
       alphabet CDATA #IMPLIED
>
<!ELEMENT break EMPTY>
<!ATTLIST break
       time CDATA #IMPLIED
       strength  (none | x-weak | weak  | medium | strong | x-strong) "medium"
>
<!ELEMENT mark EMPTY>
<!ATTLIST mark
       name CDATA #REQUIRED
>

<!ELEMENT lexicon EMPTY>
<!ATTLIST lexicon
    uri %uri; #REQUIRED
    type CDATA #IMPLIED
>

<!ELEMENT metadata ANY>

<!ELEMENT meta EMPTY>
<!ATTLIST meta
       name NMTOKEN #IMPLIED
       content CDATA #REQUIRED
       http-equiv NMTOKEN #IMPLIED
>
```

```
<!-- TC-STAR Module interface specification -->


<!-- Token, word, POS, transcription and phonemes (BEGIN) -->
<!ELEMENT TOKEN (WORD+)>
<!ATTLIST TOKEN token CDATA "not_specified">

<!ELEMENT WORD (POS, PHONETIC, PHON*)>
<!ATTLIST WORD word CDATA "not_specified">

<!--ELEMENT POS (%pos;)-->
<!ELEMENT POS (%pos;)>

<!ELEMENT PHON (frequency, energy, voice-quality, syllable)*>
<!ATTLIST PHON  phoneme CDATA "not_specified"
                duration CDATA "not_specified"
>
<!-- Token, word, transcription and phonemes (END) -->

<!-- Voice Quality (BEGIN) -->
<!ELEMENT voice-quality (laryngeal, tension, source)*>

<!ELEMENT laryngeal EMPTY>
<!ATTLIST laryngeal     modal CDATA "not_specified"
                        falsetto CDATA "not_specified"
                        whisper CDATA "not_specified"
                        creak CDATA "not_specified"
                        harshness CDATA "not_specified"
                        breathiness CDATA "not_specified"
>

<!ELEMENT tension EMPTY>
<!ATTLIST tension mode (%tension;) "not_specified">

<!ELEMENT source EMPTY>
<!ATTLIST source        EE CDATA "not_specified"
                        OQ CDATA "not_specified"
                        AS CDATA "not_specified"
                        RA CDATA "not_specified"
                        RG CDATA "not_specified"
                        RK CDATA "not_specified"
>
<!-- Voice Quality (END) -->

<!-- Prosody elements (BEGIN) -->
<!ELEMENT energy (pair)+>

<!ELEMENT frequency (pair)+>

<!ELEMENT pair EMPTY>
<!ATTLIST pair  time CDATA #REQUIRED
                value CDATA #REQUIRED
>
<!-- Prosody elements (END) -->

<!-- Syllable (BEGIN) -->
<!ELEMENT syllable (break*)>
<!ATTLIST syllable last-syllable (%boolean;) "not_specified"
                accent (%accent;) "not_specified"
>
<!-- Syllable (END) -->
```

4.7 LC-STAR DTD

This DTD (copied from [Mal 04]) is the base of the POS tagging and the phonetic transcription in the TC-STAR definition of the inter-module interfaces. It was created during the LC-STAR project in order to provide a formal definition of the lexica needed for the different languages supported there, that at the time of writing are:

- Catalan,
- Finnish,
- German,
- Greek,
- Hebrew,
- Italian,
- Mandarin Chinese,
- Russian,
- Slovenian,
- Spanish,
- Standard Arabic,
- Turkish,
- US-English.

See [Mal 04] for a more detailed explanation.

```
<?xml version="1.0" encoding="UTF-16"?>
<!-- Language-independent specification of contents of lexica -->
<!-- Associated to version 1.1 of D2 document. Sep 23, 2003 -->

<!-- Entity Declarations (BEGIN) -->
<!ENTITY % ns "not_specified">
<!ENTITY % pos "NOM | ADJ | DET | NUM | VER | AUX | PRO |
                ART | ADV | CON | ADP | INT | PAR | PRE |
                ONO | MEW | AUW | IDI |  PUN | ABB | LET">
<!ENTITY % subdomain "0.1.1. | 0.1.2. | 0.1.3. | 0.2. | 1.1.1. |
                      1.1.2. | 1.1.3. | 1.1.4. | 1.2.1. | 1.2.2. | 1.3. |
                      1.4. | 1.5.1. | 1.5.2. | 1.6. | 2.1.1. | 2.1.2. |
                      2.1.3. | 2.1.4. | 2.2.1. | 2.2.2. | 2.2.3. |
                      3.1.1. | 3.1.2. | 3.1.3. | 3.1.4. | 3.1.5. | 4.1.1. |
                      4.1.2. | 4.1.3. | 4.1.4. | 4.1.5. | 4.1.6. | 5.1.1. |
                      5.1.2. | 5.1.3. | 5.1.4. | 5.1.5. | 5.2. |
                      6.1.1. | 6.1.2. | 6.1.3. | 6.1.4. | 6.1.5. |
                      6.2.1. | 6.2.2. | 6.2.3. | 6.2.4. | 6.2.5. |
                      6.2.6.">
<!ENTITY % class_noun "common | PER | GEO | COU |
                         CIT | STR | COM | BRA | TOU | HLD">
<!ENTITY % number "singular | plural | invariant |
                                  dual | %ns;">
<!ENTITY % number_adj "%number; | general">
<!ENTITY % gender "masculine | feminine | neuter | invariant | %ns;">
<!ENTITY % gender_adj "%gender; | general">
<!ENTITY % person "1 | 2 | 3| invariant | %ns;">
<!ENTITY % person_ver "%person; | not_3">
<!ENTITY % case "nominative | genitive | partitive |
                 genitive_partitional | essive | translative |
                 inessive | elative | illative | adessive |
                 ablative | allative | abessive | instructive |
                 comitative | accusative | vocative | dative |
                 locative | instrumentative | equative |
                  prepositional | indeclinable | invariant | %ns;">
```

```
<!ENTITY % type_noun "animated | not_animated | possessive |
                      construct_case | agent | ness | zero |
                      past_participle | future_participle |
                      infinitive | feel_like | not_state |
                      not_able_state | act_of | diminutive | %ns;">
<!ENTITY % appreciative "diminutive | augmentative |
                         pejorative | %ns;">
<!ENTITY % possessive_agreement "none | SG1 | SG2 | SG3 |PL1 | PL2 |
                                 PL3 | %ns;">
<!ENTITY % degree "positive | comparative | superlative | %ns;">
<!ENTITY % form "full | concise | %ns;">
<!ENTITY % type_adj "qualitative | relative | possessive | with | without |
                     fit_for | in_between | agent | past_participle |
                     future_participle | present_participle | construct_case |
                     feel_like | related | just_like | zero | %ns;">
<!ENTITY % tense "present | imperfect | past | narrative | pluperfect |
                  aorist | future | narrative_past | future_past |
                  future_narrative | past_past | narrative_narrative |
                  imperative | aorist_passive | %ns;">
<!ENTITY % mood "indicative | subjunctive | conditional | optative |
                 imperative | infinitive | infinitiveI |
                 infinitiveII | infinitiveIII | infinitiveIV |
                 necessitative | desirative | participle |
                 adverbial_participle |participleI | participleII |
                 gerund | potential | progressive | progressiveII |
                 participle_present | participle_perfect | finite |
                 progressive | %ns;">
<!ENTITY % flag "yes | no | %ns;">

<!-- Entity Declarations (END) -->

<!ELEMENT LEXICA (ENTRYGROUP)+>
<!ATTLIST LEXICA xml:lang NMTOKEN  #IMPLIED>
<!ELEMENT ENTRYGROUP (ALT_SPEL*, (ENTRY | ENTRY_COMP | ABB)+)>
<!ATTLIST ENTRYGROUP orthography CDATA #REQUIRED
                     xml:lang NMTOKEN  #IMPLIED >
<!ELEMENT ALT_SPEL (#PCDATA)>
<!ELEMENT ENTRY_COMP (PHONETIC, LEMMA*, ENTRY_EL, ENTRY_EL, ENTRY_EL*, APP?)>
<!ELEMENT PHONETIC (#PCDATA)>
<!ELEMENT ENTRY_EL (%pos;)>
<!ATTLIST ENTRY_EL orthography CDATA #REQUIRED>
<!ELEMENT ABB (EXP)+>
<!ELEMENT EXP (ENTRY_COMP | ENTRY)>
<!ATTLIST EXP expansion CDATA #IMPLIED>
<!ELEMENT ENTRY ((%pos;), LEMMA, PHONETIC, APP?)>
<!ELEMENT LEMMA (#PCDATA)>
<!ELEMENT APP (SBD+)>
<!ELEMENT SBD EMPTY>
<!ATTLIST SBD
        type (%subdomain;) #REQUIRED
    entries CDATA #REQUIRED>

<!-- POS DEFINITION BEGIN -->
<!ELEMENT NOM EMPTY>
<!ATTLIST NOM
        class (%class_noun;) #REQUIRED
        number (%number;) "not_specified"
        gender (%gender;) "not_specified"
        case (%case;) "not_specified"
        type (%type_noun;) "not_specified"
        appreciative (%appreciative;) "not_specified"
        poss_agreem (%possessive_agreement;) "not_specified">
<!ELEMENT ADJ EMPTY>
<!ATTLIST ADJ
        number (%number_adj;) "not_specified"
```

```
        gender (%gender_adj;) "not_specified"
        case (%case;) "not_specified"
        degree (%degree;) "not_specified"
        form (%form;) "not_specified"
        type (%type_adj;) "not_specified"
        appreciative (%appreciative;) "not_specified"
        poss_agreem (%possessive_agreement;) "not_specified">
<!ELEMENT DET EMPTY>
<!ATTLIST DET
        number (%number;) "not_specified"
        gender (%gender_adj;) "not_specified"
        person (%person;) "not_specified"
        case (%case;) "not_specified"
        form (%form;) "not_specified"
        type (possessive | demonstrative | indefinite | interrogative |
            exclamative | relative | pronominal | definite | negative |
            definite_article | attributive | %ns;) "not_specified"
        degree (%degree;) "not_specified">
<!ELEMENT NUM EMPTY>
<!ATTLIST NUM
        number (%number;) "not_specified"
        gender (%gender;) "not_specified"
        case (%case;) "not_specified"
        type (ordinal | cardinal | multiplicative | collective |
              percentage | real | range | ratio | distributive | relative |
              time | construct_case | indefinite | %ns;) "not_specified">
<!ELEMENT VER EMPTY>
<!ATTLIST VER
        number (%number;) "not_specified"
        gender (%gender;) "not_specified"
        person (%person_ver;) "not_specified"
        case (%case;) "not_specified"
        mood (%mood;) "not_specified"
        tense (%tense;) "not_specified"
        voice (active | passive | reflexive | pronominal | %ns;)
              "not_specified"
        polarity (positive | negative | %ns;) "not_specified"
        aspect (perfect | imperfect | progressiveI |
                progressiveII | aorist | %ns;) "not_specified"
        form (%form;) "not_specified"
        copula (%flag;) "not_specified"
        type (causative | reflexive | passive | reciprocal_collective |
              become | acquire | able | repeat | hastily | ever_since |
              almost | stay | start | continue | zero | %ns;) "not_specified">
<!ELEMENT AUX EMPTY>
<!ATTLIST AUX
        number (%number;) "not_specified"
        gender (%gender;) "not_specified"
        person (%person_ver;) "not_specified"
        case (%case;) "not_specified"
        tense (%tense;) "not_specified"
        mood (%mood;) "not_specified"
        voice (active | passive | reflexive | %ns;) "not_specified"
        polarity (positive | negative | %ns;) "not_specified"
        aspect (perfect | imperfect | %ns;) "not_specified"
        form (%form;) "not_specified"
        type (finite | modal | %ns;) "not_specified">
<!ELEMENT PRO EMPTY>
<!ATTLIST PRO
        number (%number;) "not_specified"
        poss_agreem (%possessive_agreement;) "not_specified"
        gender (%gender; | indeterminate) "not_specified"
        person (%person;) "not_specified"
        case (%case; | oblique) "not_specified"
        type (personal | demonstrative | reflexive | indefinite |
```

```
              interrogative | reciprocal | relative | possessive | definite |
              exclamative | quantifying | negative | %ns;) "not_specified"
        politeness (%flag;) "not_specified">
<!ELEMENT ART EMPTY>
<!ATTLIST ART
        number (%number;) "not_specified"
        gender (%gender;) "not_specified"
        case (%case;) "not_specified"
        type (definite | indefinite | partitive | %ns;) "not_specified">
<!ELEMENT ADV EMPTY>
<!ATTLIST ADV
        degree (%degree;) "not_specified"
        type (time | place | after_doing_so | since | when | by_doing_so |
              while | as_if | without_having_done_so | ly | adamantly |
              without_being_able_to_have_done_so | as_long_as |
            since_doing_so | manner | %ns;) "not_specified">
<!ELEMENT CON EMPTY>
<!ATTLIST CON
        type (coordinating | subordinating | %ns;) "not_specified">
<!ELEMENT ADP EMPTY>
<!ATTLIST ADP
        number (%number;) "not_specified"
        gender (%gender;) "not_specified"
        person (%person;) "not_specified"
        type (simple | articulated | possessive | %ns;) "not_specified">
<!ELEMENT INT EMPTY>
<!ELEMENT PAR EMPTY>
<!ATTLIST PAR
        number (%number;) "not_specified"
        person (%person;) "not_specified"
        tense (present | past | narrative | %ns;) "not_specified"
        mood (conditional | %ns;) "not_specified"
        copula (yes | %ns;) "not_specified">
<!ELEMENT PRE EMPTY>
<!ELEMENT ONO EMPTY>
<!ELEMENT MEW EMPTY>
<!ELEMENT AUW EMPTY>
<!ELEMENT IDI EMPTY>
<!ELEMENT PUN EMPTY>
<!ELEMENT LET EMPTY>
<!-- POS DEFINITION END -->
```

4.8 TC-STAR XML Examples

Below you will find a short example of an input text in SSML format, and its correspondent TC-STAR XML documents at the input of the prosody and synthesis modules. As you will see, the SSML text contains a mark to specify a prosodic change in the pronunciation of one word. Although the text processing module does not know anything about how the prosody is specified, it maintains the mark so the prosodic module can use the information (resulting in longer durations of the affected phonemes).

Please note that the examples below are merely informative and are intended to illustrate the usage and position of the different tags available to the developers. In particular, voice quality tags have been randomly chosen to show several valid combinations.

4.8.1 SSML input

```
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE speak PUBLIC "-//W3C//DTD SYNTHESIS 1.0//EN"
                  "http://www.w3.org/TR/speech-synthesis/synthesis.dtd">
<speak version="1.0"
  xmlns="http://www.w3.org/2001/10/synthesis"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
```

```
  xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
  http://www.w3.org/TR/speech-synthesis/synthesis.xsd"
       xml:lang="es">

    <s>Cuerpo <prosody rate="-20%">gaseiforme</prosody> que sin embargo ofrece
resistencia.</s>

</speak>
```

4.8.2 Prosody module input

```
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE tts PUBLIC "TC-STAR"
                  "TC-STAR.dtd">
<tts xml:lang="es">
  <s>
    <TOKEN token="Cuerpo">
      <WORD word="Cuerpo">

        <POS>
          <NOM class="common" number="singular" gender="masculine"/>
        </POS>

        <PHONETIC>kw'eR - po</PHONETIC>

      </WORD>
    </TOKEN>

    <prosody rate="-20%">
      <TOKEN token="gaseiforme">
        <WORD word="gaseiforme">
          <POS>
            <ADJ number="singular" gender="neuter" type="qualitative"/>
          </POS>
          <PHONETIC>Ga-sej-f'or-me</PHONETIC>
        </WORD>
      </TOKEN>
    </prosody>

    <TOKEN token="que">
      <WORD word="que">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
      </WORD>
    </TOKEN>

    <TOKEN token="sin embargo">
      <WORD word="sin">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
      </WORD>
      <WORD word="embargo">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
      </WORD>
    </TOKEN>

    <TOKEN token="ofrece">
      <WORD word="ofrece">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
      </WORD>
    </TOKEN>
```

```
    <TOKEN token="resistencia.">
      <WORD word="resistencia">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
      </WORD>
      <WORD word=".">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
      </WORD>
    </TOKEN>

  </s>
</tts>
```

4.8.3 Synthesis module input

```
<?xml version="1.0" encoding="UTF-16"?>
<!DOCTYPE tts PUBLIC "TC-STAR"
                  "TC-STAR.dtd">
<tts xml:lang="es">
  <s>
    <TOKEN token="Cuerpo">
      <WORD word="Cuerpo">

        <POS>
          <NOM class="common" number="singular" gender="masculine"/>
        </POS>

        <PHONETIC>kw'eR - po</PHONETIC>

        <PHON phoneme="k" duration="5">
          <frequency>
            <pair time="0" value="110"/>
            <pair time="5" value="116"/>
          </frequency>
          <energy>
            <pair time="0" value="24"/>
            <pair time="4" value="22"/>
          </energy>
          <voice-quality>
            <laryngeal falsetto="75" harshness="25"/>
          </voice-quality>
          <syllable last-syllable="false" accent="primary">
            <break time="1" strength="weak"/>
          </syllable>
        </PHON>

        <PHON phoneme="w" duration="4">
          <frequency>
            <pair time="0" value="110"/>
            <pair time="4" value="115"/>
          </frequency>
          <energy>
            <pair time="0" value="24"/>
            <pair time="4" value="22"/>
          </energy>
          <voice-quality>
            <laryngeal falsetto="72" creak="3" harshness="25"/>
          </voice-quality>
        </PHON>

        <PHON phoneme="e" duration="6">
          <frequency>
            <pair time="0" value="110"/>
            <pair time="4" value="115"/>
```

```
            <pair time="6" value="117"/>
          </frequency>
          <energy>
            <pair time="0" value="24"/>
            <pair time="2" value="22"/>
            <pair time="6" value="21"/>
          </energy>
          <voice-quality>
            <laryngeal falsetto="70" creak="7" harshness="23"/>
          </voice-quality>
        </PHON>

        <PHON phoneme="p" duration="2">
          <frequency>
            <pair time="0" value="110"/>
            <pair time="2" value="117"/>
          </frequency>
          <energy>
            <pair time="0" value="24"/>
            <pair time="2" value="22"/>
          </energy>
          <voice-quality>
            <laryngeal falsetto="70" creak="7" harshness="23"/>
          </voice-quality>
          <syllable last-syllable="true">
            <break time="4" strength="strong"/>
          </syllable>
        </PHON>

        <PHON phoneme="o" duration="5">
          <frequency>
            <pair time="0" value="110"/>
            <pair time="2" value="117"/>
            <pair time="5" value="119"/>
          </frequency>
          <energy>
            <pair time="0" value="24"/>
            <pair time="2.5" value="22"/>
            <pair time="5" value="25"/>
          </energy>
          <voice-quality>
            <laryngeal falsetto="80" harshness="20"/>
          </voice-quality>
        </PHON>
      </WORD>
    </TOKEN>

    <prosody rate="-20%">
      <TOKEN token="gaseiforme">
        <WORD word="gaseiforme">
          <POS>
            <ADJ number="singular" gender="neuter" type="qualitative"/>
          </POS>
          <PHONETIC>Ga-sej-f'or-me</PHONETIC>

          <PHON phoneme="G" duration="5">
            <frequency>
              <pair time="0" value="110"/>
              <pair time="5" value="116"/>
            </frequency>
            <energy>
              <pair time="0" value="24"/>
              <pair time="4" value="22"/>
            </energy>
            <voice-quality>
```

```
            <laryngeal modal="78" harshness="22"/>
          </voice-quality>
          <syllable last-syllable="false" accent="secundary">
            <break time="1" strength="x-weak"/>
          </syllable>
        </PHON>

        <PHON phoneme="a" duration="5">
          <frequency>
            <pair time="0" value="110"/>
            <pair time="5" value="116"/>
          </frequency>
          <energy>
            <pair time="0" value="24"/>
            <pair time="4" value="22"/>
          </energy>
          <voice-quality>
            <laryngeal modal="85" harshness="10" breathiness="5"/>
          </voice-quality>
        </PHON>

        <PHON phoneme="s" duration="2">
          <frequency>
            <pair time="0" value="115"/>
            <pair time="2" value="120"/>
          </frequency>
          <energy>
            <pair time="0" value="22"/>
            <pair time="2" value="21"/>
          </energy>
          <voice-quality>
            <laryngeal modal="85" harshness="10" breathiness="5"/>
            <tension mode="metallic"/>
          </voice-quality>
          <syllable last-syllable="false">
          </syllable>
        </PHON>

        <PHON phoneme="e" duration="3">
          <frequency>
            <pair time="0" value="115"/>
            <pair time="2" value="120"/>
            <pair time="3" value="122"/>
          </frequency>
          <energy>
            <pair time="0" value="22"/>
            <pair time="3" value="21"/>
          </energy>
          <voice-quality>
            <tension mode="strident"/>
          </voice-quality>
        </PHON>

        <PHON phoneme="j" duration="3">
          <frequency>
            <pair time="0" value="115"/>
            <pair time="2" value="120"/>
            <pair time="3" value="122"/>
          </frequency>
          <energy>
            <pair time="0" value="22"/>
            <pair time="3" value="21"/>
          </energy>
          <voice-quality>
            <tension mode="strident"/>
```

```
            <source EE="-15" OQ="70" AS="10" RA="15" RG="10" RK="11"/>
          </voice-quality>
        </PHON>

        <PHON phoneme="f" duration="2">
          <frequency>
            <pair time="0" value="115"/>
            <pair time="2" value="122"/>
          </frequency>
          <energy>
            <pair time="0" value="22"/>
            <pair time="2" value="21"/>
          </energy>
          <voice-quality>
            <laryngeal modal="100"/>
          </voice-quality>
          <syllable last-syllable="false" accent="primary">
            <break time="2" strength="weak"/>
          </syllable>
        </PHON>

        <PHON phoneme="o" duration="3">
          <frequency>
            <pair time="0" value="115"/>
            <pair time="1" value="117"/>
            <pair time="2" value="116"/>
            <pair time="3" value="118"/>
          </frequency>
          <energy>
            <pair time="1" value="22"/>
            <pair time="3" value="23"/>
          </energy>
          <voice-quality>
            <laryngeal modal="100"/>
          </voice-quality>
        </PHON>

        <PHON phoneme="R" duration="1">
          <frequency>
            <pair time="0" value="115"/>
            <pair time="1" value="117"/>
          </frequency>
          <energy>
            <pair time="1" value="22"/>
          </energy>
          <voice-quality>
            <laryngeal modal="100"/>
          </voice-quality>
        </PHON>

        <PHON phoneme="m" duration="2">
          <frequency>
            <pair time="0" value="115"/>
            <pair time="2" value="122"/>
          </frequency>
          <energy>
            <pair time="0" value="22"/>
            <pair time="2" value="21"/>
          </energy>
          <voice-quality>
            <laryngeal modal="100"/>
          </voice-quality>
          <syllable last-syllable="true" accent="not_specified">
            <break time="0.5" strength="weak"/>
          </syllable>
```

```
          </PHON>

          <PHON phoneme="e" duration="3">
            <frequency>
              <pair time="0" value="115"/>
              <pair time="1" value="117"/>
              <pair time="2" value="116"/>
              <pair time="3" value="118"/>
            </frequency>
            <energy>
              <pair time="1" value="22"/>
              <pair time="3" value="23"/>
            </energy>
            <voice-quality>
              <laryngeal modal="100"/>
            </voice-quality>
          </PHON>
        </WORD>
      </TOKEN>
    </prosody>

    <TOKEN token="que">
      <WORD word="que">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
       <PHON> ... </PHON>
       <PHON> ... </PHON>
      </WORD>
    </TOKEN>

    <TOKEN token="sin embargo">
      <WORD word="sin">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
       <PHON> ... </PHON>
       <PHON> ... </PHON>
      </WORD>
      <WORD word="embargo">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
       <PHON> ... </PHON>
       <PHON> ... </PHON>
      </WORD>
    </TOKEN>

    <TOKEN token="ofrece">
      <WORD word="ofrece">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
       <PHON> ... </PHON>
       <PHON> ... </PHON>
      </WORD>
    </TOKEN>

    <TOKEN token="resistencia.">
      <WORD word="resistencia">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
       <PHON> ... </PHON>
       <PHON> ... </PHON>
      </WORD>
      <WORD word=".">
        <POS> ... </POS>
        <PHONETIC> ... </PHONETIC>
      </WORD>
```

```
    </TOKEN>

  </s>
</tts>
```

4.9 References

[Mal 04] G. Maltese and C. Montecchio, “General and language-specific specification of contents of lexica in 13 languages,” LC-STAR Deliverable, May 2004. [Online]. Available: http://www.lc-star.com/WP2_deliverable_D2_v2.1.doc =0pt

[Bur 04] D. C. Burnett, M. R. Walker, and A. Hunt, “Speech synthesis markup language (SSML) version 1.0,” W3C Recommendation, Sept. 2004. [Online]. Available: http://www.w3.org/TR/speech-synthesis/ =0pt

[Lav 80] J. Laver, *The Phonetic Description of Voice Quality*.1em plus 0.5em minus 0.4emCambridge University Press, 1980.

[Kel --] E. Keller, *Lecture Notes in Computer Science*.1em plus 0.5em minus 0.4emSpringer Verlag, ch. The Analysis of Voice Quality in Speech Processing, to be published.

[Gob 03] C. Gobl, “The voice source in speech communication,” Ph.D. dissertation, Department of Speech, Music and Hearing, KTH, Stockholm, 2003.

[Spr 99] Richard Sproat, Alan Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. “Normalization of non-standard words: WS'99 final report”. Technical Report, September 1999.