ECESS Inter-Module Interface Specification for Speech Synthesis

Javier Pérez [⋆], Antonio Bonafonte [⋆], Horst-Udo Hain [⊥], Eric Keller [†], Stefan Breuer [‡], Jilei Tian [∗]

[⋆] TALP, Universitat Politècnica de Catalunya, Barcelona, Spain
{javierp,antonio}@gps.tsc.upc.edu
[⊥] Siemens AG, Corporate Technology, Munich, Germany
horst-udo.hain@siemens.com
[†] LAIP - IMM - Lettres, Université de Lausanne, Switzerland
eric.keller@unil.ch
[‡] IfK, University of Bonn, Germany,
breuer@ikp.uni-bonn.de
[∗] Multimedia Technologies Laboratory, Nokia Research Center, Tampere, Finland
jilei.tian@nokia.com

Abstract

The newly founded European Centre of Excellence for Speech Synthesis (ECESS) (ECESS, 2004) is an initiative to promote the development of the European research area (ERA) in the field of Language Technology. ECESS focuses on the great challenge of high-quality speech synthesis which is of crucial importance for future spoken-language technologies. The main goals of ECESS are to achieve the critical mass needed to promote progress in TTS technology substantially, to integrate basic research know-how related to speech synthesis and to attract public and private funding. To this end, a common system architecture based on exchangeable modules supplied by the ECESS members is to be established. The XML-based interface that connects these modules is the topic of this paper.

1. Introduction

One of the objectives of ECESS is to design a common system architecture for speech synthesis based on well-defined modules and interfaces. The modules are interchangeable and are evaluated using a common set of evaluation criteria. Different partners supply these modules and evaluate the required language resources using a common specification. Hence, each institution focuses R&D on (at least) one module, providing it license-free for research use to the other partners of the consortium. The infrastructure of TC-STAR is used to periodically evaluate the system and the individual modules (TC-STAR, 2004).

The three main modules in the ECESS approach follow a commonly employed approach to the text-to-speech task: **symbolic pre-processing** (performs the tokenization, POS tagging and phonetic transcription of the input text), **prosody generation** (the system uses acoustic prosody: silences, duration, energy and fundamental frequency of the phones) and **acoustic synthesis** (voice generation according to the prosodic specification). Since existing components will have to be adapted to the modular architecture used by the ECESS consortium, we have created a formal definition of the inter-module interfaces using XML and a DTD to facilitate the integration into the common framework. We chose XML since several synthesis systems are already capable of processing and generating some XML-based languages (VoiceXML, SSML or particular implementations). The advantage of an XML-based interface is that existing libraries and software can be used for the generation, validation and parsing of data, thus ensuring a fast and flexible interface implementation and re-definition.

Our interface definition formally describes the communication format between the text processing and the prosody generation modules, and between the prosody generation and acoustic synthesis modules. Since each module performs a complementary task, only one DTD is necessary. Each module adds information to the corresponding part of the XML document while maintaining the information previously added by any other module. The basis for our interface is the Speech Synthesis Markup Language (SSML). As SSML is an XML specification that focuses on describing TTS input rather than representing the phonetic and acoustic details needed for speech synthesis (for a detailed discussion, see (Schröder and Breuer, 2004)), we had to extend the format in a number of ways.

2. Symbolic pre-processing

The first step in a TTS system is to analyze the input text and to transform it into a linguistic representation containing all the necessary information needed in the subsequent steps of the synthesis. The main modules of the symbolic pre-processing step are: tokenizer, morphology analyzer, POS tagger, and grapheme-to-phoneme converter. In the ECESS synthesis system, the processing module will be designed in a way to encounter multilingual and polyglot aspects of text processing as much as possible. Therefore, all the language dependent resources will be separated from the language independent text processing engine. The symbolic pre-processing module performs the tokenization, POS tagging and phonetic transcription of the input text, identifying numbers, acronyms, abbreviations and other special symbols, and expands them into full text form.

Each XML element of type *token* consists of zero or more *word* elements, each of them having an associated *transcription* and *POS* element. Phonetic transcription information is to be coded for each word in the way that the word is spoken in isolation. We use the SAMPA phonetic alphabet with syllable boundary marker (-), stress marker (¨) and tone markers for tonal languages. POS coding is

partly based on the formal definition specified by the LC-STAR (Maltese and Montecchio, 2004) project, among others: NOM (name), ADJ (adjective), ADV (adverb), PRE (preposition), DET (determinant). Figure 1 shows a partial example of the output of this module for the Spanish language. The data was extracted from the output of the module during the first evaluation campaign of TC-STAR.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE tts SYSTEM "TC-STAR.dtd">
<tts xml:lang="es">
<p>
<s>
<TOKEN token="En">
  <WORD word="En">
    <POS>
      <ADP/>
    </POS>
    <PHONETIC>e n</PHONETIC>
  </WORD>
</TOKEN>
<TOKEN token="ningún">
  <WORD word="ningún">
    <POS>
      <ADJ/>
    </POS>
    <PHONETIC>n i N - g " u n</PHONETIC>
  </WORD>
</TOKEN>
<TOKEN token="momento">
  <WORD word="momento">
    <POS>
      <NOM class="common"/>
    </POS>
    <PHONETIC>m o - m " e n - t o</PHONETIC>
  </WORD>
</TOKEN>
</s>
</p>
</tts>
```

Figure 1: Spanish sample output from the symbolic preprocessing module.

3. Prosody generation

Prosody is the set of speech features that allows the same phonetic sounds to be uttered in different ways, containing linguistic, sociolinguistic and expressive information. At the linguistic level, prosody indicates the sentence type and structures of the utterances, producing chunks of words for some syntactic, semantic or even pragmatic reason. At the sociolinguistic level, prosody provides information about the speaker, including his social or cultural dialect. And prosody is fundamental for expressing intentions and attitudes of the speaker about the linguistic information. Therefore, prosody plays a fundamental role in eliciting the meaning, attitude and intention and producing natural speech.

Our approach to the interface between the prosody and synthesis module focuses on the necessary acoustics features. These features include intonation (tone, pitch contour), speech rate, segment duration, phrase break, stress level and voice quality. The basic unit of analysis for the prosody module will be the phones (this is not the case for Asian languages, where the basic unit is the syllable, as will be explained in section 5.. Figures 2 and 3 contain two working examples of the prosody module output for Spanish and Chinese respectively.

3.1. Duration, frequency and intensity

The prosody generation module will associate with each word a list of corresponding phones. This does not necessarily need to be equal to the phonetic transcription itself as given by the text analysis module, since vowel assimilation, diphthong creation, speech rate, pauses and other phenomena may have to be considered. Each phone will have a reference *duration* expressed in *milliseconds*, a fundamental frequency contour and an energy or intensity contour. Each prosody generation module for producing these contours specifies the sampling rate (resolution) used; it is then the task of the synthesis module to use this information appropriately. As a particular example, curves sampled every 5 milliseconds are used during the TC-STAR evaluation campaign.

3.2. Syllabic information

The prosody module is required to mark the *beginning of a syllable*, and whether the syllable is the *last syllable* of a word . In our approach, this information is included at the phone level, since this methodology allows for the disassociation of words and syllables. Thus, phones of different words can be easily associated with the same syllable (this is particularly useful in case of the linking phenomena, for instance).

In order to label the break index tier of the last syllable in a word, we will follow the guidelines set by SSML (Burnett et al., 2004), where five categories are defined: `none, x-weak, weak, medium, strong` and `x-strong`. Each category refers to the *strength* of the break, and this information will be coded within the first phone of that syllable.

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).

3.3. Voice quality

Voice quality and how to use it in speech synthesis algorithms is a topic in active research. Speakers can generally be identified by distinct speech characteristics that reflect psychological dimensions reaching from the distinction of personality types, via the communication of affect and emotion, to the transmission of delicate nuances in conversational exchanges, and/or sociological markers which identifies their position in a social hierarchy or their membership in a group; in some languages, voice quality markers have been integrated into linguistic distinctiveness markers.

Following Keller's review of the field (Keller, 2005), we specified the definition of voice quality in terms of (a) voice properties resulting from either laryngeal or tension-related articulatory conditions, and of (b) manipulations of the glottal source waveform. Consequently, the following

articulatorily defined voice-forms are available and can be combined with each other: (a) laryngeal: modal, falsetto, whisper, creak, harshness and breathiness; (b) tensing or laxing of the entire vocal tract musculature: tense, sharp, shrill, metallic, strident, lax, soft, dull, guttural or mellow. Only one type of voice can be specified using this definition.

For researchers working with the glottal excitation model, the following glottal source related parameters have proved the most useful, and were thus included in our proposal: excitation energy (EE), open quotient (OQ), aspiration noise (AS), sharpness of glottal closure (RA), glottal asymmetry (RK) and glottal frequency (RG). EE and AS are power measures and will be expressed in dB. OQ, RA, RG and RK are expressed as fractions of the pitch period (in %). Work in this area is already started and preliminary results of analysis/synthesis procedures using these glottal source parameters have been reported in (Pérez and Bonafonte, 2005).

In general, voice quality information is considered an optional specification, since not all synthesis procedures require this knowledge.

4. Acoustic synthesis

The objective for the acoustic module is to generate the speech waveform using the prosodic information generated during the stages already described. The intended parameters and contours (energy of the phones, fundamental frequency, segment durations, phonetic coarticulation, etc.) should be matched in the best possible manner. No restriction is placed by the consortium on the type of technique used to perform the synthesis. Many state-of-the-art systems are based on unit selection and signal concatenation techniques. In this case, a database of predefined units is used, from which appropriate units are selected and concatenated following the requirements by the previous modules. This usually requires some prosodic manipulation to attain a closer match to the target parameters. However, other techniques can be used, for instance those based on the parameterization of the speech signal using relevant mathematical models. These techniques are normally able to provide signal manipulation of greater quality with smaller databases, with the drawback of degrading signal naturalness.

5. Chinese Adaptation

Asian languages are different from western languages with respect to a few perspectives: first, many Asian languages are tonal languages, e.g. Chinese including different spoken variants called dialects, as well as Vietnamese, Thai, are all languages in which tones play an important semantic role; second, many Asian languages are syllabic in nature, e.g. Chinese, Vietnamese, Thai, Hindi; third, many Asian languages (e.g. Chinese, Thai, Vietnamese, Japanese) do not have a proper boundary for words. In syllabic languages, syllables or sub-syllabic structures like initials and finals could be very natural units for the TTS systems (this is the case for Chinese, Vietnamese, Hindi, etc). Phones can be good units for western languages, but are not optimal for Asian languages. We should respect the characteristics of the language to be processed and give more flexibility in the XML interface design to better support different languages. These peculiarities cause special challenges for TTS systems, and consequently, require special treatment of the XML interface design.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE tts SYSTEM "TC-STAR.dtd">
<tts xml:lang="es">
<p>
<s>
<TOKEN token="En">
  <WORD word="En">
   <POS>
    <ADP/>
   </POS>
   <PHONETIC>e n</PHONETIC>
   <PHON duration="74.5" phoneme="e">
      <frequency>
        <pair time="5"  value="200.3"/>
        <pair time="20" value="197.2"/>
        <pair time="35" value="194.6"/>
        <pair time="50" value="192.4"/>
        <pair time="65" value="190.5"/>
      </frequency>
      <energy>
        <pair time="0"  value="70.72"/>
      </energy>
      <syllable last-syllable="true"/>
   </PHON>

   <PHON duration="67.9" phoneme="n">
      <frequency>
        <pair time="5"  value="188.7"/>
        <pair time="20" value="187.6"/>
        <pair time="35" value="186.8"/>
        <pair time="50" value="186.3"/>
        <pair time="65" value="186.0"/>
      </frequency>
      <energy>
        <pair time="0"  value="69.19"/>
      </energy>
   </PHON>
  </WORD>
</TOKEN>
</s>
</p>
</tts>
```

Figure 2: Spanish sample output from the prosody generation module.

The interface specification of ECESS has been adapted to Mandarin Chinese (Tian et al., 2005), which, being a tonal and syllable-based language, requires modifications with respect to the inclusion of this information. In the original specification, the basic unit of text analysis is the word, but in Mandarin it is the syllable. The Mandarin specification indicates the prosodic boundary level in the interface between text analysis and prosody generation, such as sentence, phrase, and word boundary, which are quite important for prosody prediction. All the prosodic information (pitch, energy, duration and breaks) is given under the newly proposed syllable element in the form of (time,

value) pairs.
Chinese language specifics must be taken into account when designing an XML interface. For Chinese, it is natural to use tonal syllables as the basic unit of a TTS system. Word segmentation is a very crucial issue in Chinese which does not have word boundaries. Thus the ”word” element is defined to mark the word segmentation either by automatic word segmentation or manual annotation to take a certain word segment. Its attribute is the segmented word. Inside the ”word” element, ”pos” is used for determining the pronunciation of a given word in the case where POS tagging does not work or the user forces the system to use a certain POS tag. ”Break” can be used for defining the break strength at a boundary such as character boundary, word boundary, prosodic phrase boundary, sentence boundary, etc.
For Chinese, the pitch contours play a very important role in rendering TTS speech. The same phone sequence or the same base form syllable with a different tone leads to completely different meanings. Therefore, it is recommended to enhance the descriptions on prosodic features, particularly on pitch. We describe the prosody features in (time, value) format. This approach gives the possibility to cover any prosodic needs. The element ”syllable” is introduced to define the given character. The elements ”frequency” and ”energy” are introduced to describe the prosodic features pitch and volume, in (time, value) format in order to have a better representation capability for prosodic features. Figure 3 shows an example of the XML interface used for the Mandarin TTS system.

6. Conclusions

This infrastructure has been used in the first TC-STAR Evaluation Workshop on Speech Synthesis, held in Kraków, Poland, on September 23rd 2005. Among the different tasks, the prosody modules for Spanish, English and Chinese, and the text processing components for English and Chinese, were evaluated using the formal definition presented in this article.

7. Acknowledgments

This work has been partially sponsored by the European Union under grant FP6-506738 (TC-STAR, 2004).

8. References

Daniel C. Burnett, Mark R. Walker, and Andrew Hunt. 2004. Speech synthesis markup language (SSML) version 1.0. W3C Recommendation, September. http://www.w3.org/TR/speech-synthesis/.

ECESS, 2004. *ECESS European Center of Excellence on Speech Synthesis*. http://www.ecess.org.

Eric Keller. 2005. The analysis of voice quality in speech processing. In Gérar Chollet, Anna Esposito, and Marcos Faundez-Zanuy, editors, *Lecture Notes in Computer Science*, volume 3445, pages 54–73. Springer-Verlag.

Giulio Maltese and Chiara Montecchio. 2004. General and language-specific specification of contents of lexica in 13 languages. LC-STAR Deliverable, May. http://www.lc-star.com/WP2_deliverable_D2_v2.1.doc.

```
<?xml version="1.0" encoding="UTF-8" ?>
<speak version="1.0" xml:lang="cn">
<s>
  <TOKEN token=="下午">
    <word word="下午">
    <pos>
      <NOUN />
    </pos>

    <syllable syl="下">
      <frequency>
        <pair time="0"   value="380" />
        <pair time="80"  value="363" />
        <pair time="160" value="340" />
        <pair time="240" value="301" />
      </frequency>
      <energy>
        <pair time="267" value="74" />
      </energy>
      <break strength="none" />
    </syllable>

    <syllable syl="午">
      <frequency>
        <pair time="0"   value="290" />
        <pair time="54"  value="285" />
        <pair time="108" value="285" />
        <pair time="162" value="290" />
      </frequency>
      <energy>
        <pair time="181" value="71" />
      </energy>
      <break strength="weak"  />
    </syllable>
   </word>
</TOKEN>
</s>
</speak>
```

Figure 3: Example XML output from the Mandarin prosody module.

Javier Pérez and Antonio Bonafonte. 2005. Automatic voice-source parameterization of natural speech. In *Proceedings of 9th European Conference on Speech Communication and Technology , Interspeech 2005*, Lisbon, Portugal.

Marc Schröder and Stefan Breuer. 2004. XML Representation Languages as a Way of Interconnecting TTS Modules. In *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*, Jeju, Korea.

TC-STAR, 2004. *TC-STAR, Technology and Corpora for Speech to Speech Translation*. http://www.tc-star.org.

Jilei Tian, Xia Wang, and Jani Nurminen. 2005. SSML extensions aimed to improve asian language TTS rendering. In *W3C Workshop on Internationalizing the Speech Synthesis Markup Language*, Beijing, China.