ECESS Inter-Module Interface Specification for Speech Synthesis

J. Pérez [⋆], A. Bonafonte [⋆], H. Höge [⊥],
E. Keller [†], S. Breuer [‡], J. Tian [⋆]

[⋆] TALP, Universitat Politècnica de Catalunya, Barcelona, Spain
{javierp,antonio}@gps.tsc.upc.edu

[⊥] Siemens AG, Munich, Germany
harald.hoege@siemens.com

[†] LAIP - IMM - Lettres, Université de Lausanne, Switzerland
eric.keller@unil.ch

[‡] IKP, University of Bonn, Germany,
breuer@ikp.uni-bonn.de

[⋆] Multimedia Technologies Laboratory, Nokia Research Center, Tampere, Finland
jilei.tian@nokia.com

The newly founded European Centre of Excellence for Speech Synthesis (ECESS) [1] is an initiative to promote the development 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.

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 [2] is used to periodically evaluate the system and the individual modules.

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 phonemes) and **acoustic synthesis** (voice generation according to the prosodic specification). Our DTD formally describes the interface 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 symbolic pre-processing module performs the tokenization, POS tagging and phonetic transcription of the input text. 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 [3] project, among others: NOM (name), ADJ (adjective), ADV (adverb), PRE (preposition), DET (determinant).

The prosody generation module will associate to each word a list of corresponding phonemes. This needs not necessarily be equal to the phonetic transcription itself, since vowel assimilation, diphthong creation, speech rate, pauses and other phenomena may have to be considered. Each phoneme will have a reference **duration** expressed in **milliseconds**, a fundamental frequency contour and an energy or intensity contour. Each prosody generation module producing these contours specifies the sampling rate (resolution) used; it is then the task of the synthesis module to use this information appropriately.

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 phoneme level, since this methodology allows for the disassociation of words and syllables, and phonemes of different words can be easily associated with the same syllable (particularly useful in case of the linking phenomena, for instance).

In order to label the break index tier, we will follow the guidelines set by SSML [4], 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).

Voice quality and how to use it in speech synthesis algorithms is a topic in an active research. As there is no widely accepted definition of *voice quality*, we follow the studies of Laver [5] as presented by E. Keller in [6]. Voice quality information is optional since not all synthesis procedures require this knowledge. It is possible to define voice quality in terms of different voice properties using articulatory correlates of the larynx and associated muscles. The following voice-forms are available and can be combined with each other: modal, falsetto, whisper, creak, harshness and breathiness. 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 or mellow. Only one type of voice can be specified using this definition. Another common approach is to work with the glottal excitation and the vocal tract separately. In this case, the following glottal source related parameters have proved the most useful, and thus we include them in our proposal: excitation energy, open quotient, aspiration noise, sharpness of glottal closure, glottal asymmetry and glottal frequency.

Since existing components will have to be adapted to this architecture, 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 have chosen XML since several synthesis systems are already capable of processing and

generating some XML-based languages (VoiceXML, SSML or particular implementations). Existing solutions focus on the input to the synthesis system, thus lacking the level of detail necessary to perform the inter-module communication. 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.

The interface specification of ECESS has been adapted to Mandarin Chinese, which, being a tonal, 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 break) depicted as (time, value) pair format is under newly proposed syllable element.

This infrastructure has been used in the first TC-STAR Evaluation Workshop on Speech Synthesis [2], held in Kraków, Poland, on September 23rd 2005.

References

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

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

[3] 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

[4] 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/

[5] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge University Press, 1980.

[6] E. Keller, "The analysis of voice quality in speech processing," in *Lecture Notes in Computer Science*, G. Chollet, A. Esposito, and M. Faundez-Zanuy, Eds. Springer-Verlag, 2005, vol. 3445, pp. 54–73.