Speech Synthesis Markup Languages: An Overview

Hannes Pirker (OFAI).

This is an excerpt from Pirker H., Krenn B.:"Assessment of Assessment of Markup Languages for Avatars, Multimedia and Multimodal Systems", NECA-project, Deliverable D9c, May 2002. Available online at http://www.oefai.at/NECA/publications
.

Speech Synthesis Markup Language (SSML)

The Speech Synthesis Markup Language (SSML) is a standard currently under development by W3C's Voice Browser Working Group. The actual version is "W3C Working Draft 5 April 2002". SSML is designed to provide an XML-based markup language for the generation of synthetic speech both in the WWW as well as in stand-alone synthesizers. It aims to give authors of synthesizable text the opportunity to control the output of synthesized speech with respect to pronunciation, volume, pitch, rate, etc. across different platforms capable of synthesis.
The development of SSML is strongly based on other existing speech markup languages –especially JSML (Java Speech Markup Language) and SABLE[1]. These will thus not be surveyed here.
The current draft of SSML comprises 12 elements which are grouped as follows:

Document Structure, Text Processing and Pronunciation

<speak>: The root element for SSML documents.
Attribute:

- lang : Specifies the language to be used (lang is also defined for <paragraph> and <sentence>)

<paragraph> and <sentence>: Represents the internal structure of texts

<say-as>: Gives information on the "type of text" included in this element, in order to aid the correct PRONUNCiation by specifying that a text is to be interpreted, e.g., as currency, date, address.

Attribute:

- type : Encompasses many different types, like pronunciation type ("acronym", "spell-out") numerical type ("number", "ordinal", "cardinal", "digits"), time and measure types ("date", "time", "currency", "measure", …)

<phoneme>: Provides phonetic pronunciation.
Attribute:

- alphabet: Specifies which phonetic alphabet is used (e.g., "ipa")

<sub>: When synthesizing substitutes the contained text by the one given in the "alias" attribute
Attribute:

- alias: specifies the text to be pronounced instead (e.g. _{WWW})

---

[1] Note, that in the literature on SABLE it is claimed, that SABLE is based on SSML, while in the W3C-SSML documents it is noted that SSML is based on SABLE. This confusion is due to the fact, that there exists an older – now obsolete – markup-language called "SSML", which then was replaced by SABLE. (Taylor P., Isard A.: SSML: A speech synthesis markup language, Speech Communication, (21), pp.123-133, 1997.)

Prosody and Style

<voice>: Specifies the voice to be used
Attributes:

- lang (optional language specification)
- gender (“male”, “female”, “neutral”)
- age (preferred age of the voice to speak the contained text – integer)
- name (platform-specific voice name)
- variant (indicating a preferred variant of the selected voice – integer)

<emphasis>: The contained text is to be spoken with emphasis.
Attribute:

- level: strength of emphasis (“none”, “reduced”, “moderate”, “strong”)

<break>: An empty element controlling pausing and realisation of prosodic boundaries
Attributes:

- size (strength of boundary (“none”, “small”, “medium”, “large” – optional)
- time (duration of a pause in seconds or milliseconds – optional)

<prosody>: Permit control over the prosody to be used
Attributes:

- pitch (the baseline in Hertz, a relative change or the values “default”, “low”, “medium”, “high”)
- contour (specify a concrete pitch contour)
- range (the pitch range in Hertz, a relative change or the values “default”, “low”, “medium”, “high”)
- rate (the speech rate in words per minute. A relative change or the values “default”, “slow”, “medium”, “fast”)
- duration (the time to be used to pronounce an item in seconds or milliseconds)
- volume (in a range from 0.0 to 100.0, a relative change or the values “default”, “silent”, “soft”, “medium”, “loud”)

Other Elements

<audio>: Embed an audio file, which is replayed when the element is reached
Attribute:

- src (specify the name of the audio file)

<mark> Place a marker in the text to be used either for internal reference within the SSML document or to be used externally by other documents. When speech synthesis reaches a <mark> element it issues an event with its name
Attribute:

- name (string issued as event name when mark is reached – required)

Online References:

The actual version of SSML is always to be found at: http://www.w3.org/TR/speech-synthesis

Microsoft SAPI TTS XML

Within its Speech Application Programming Interface (SAPI), Microsoft offers its own interface to speech synthesis. The current version is SAPI 5.1. Though SAPI itself is not a markup language SAPI 5.1 offers an XML based language for TTS which is explicitly inspired by SABLE (one of SSML's ancestors) but is not aiming for real compatibility with SABLE. As it has to be taken into account, that Microsoft's power on the market tends to apply some pressure on every standardization effort, we will perform a short comparison of SAPI 5.1. TTS XML with SSML.
Roughly speaking, many of the differences are only minor deviations in terminology (e.g. <ssml:say_as> and <sapi:context> are basically identical in function). Some differences arise from the fact that a functionality is expressed in terms of its own element in the one markup-language but in terms of an attribute in the other (e.g. <ssml:say_as type="spell"> and <sapi:spell> have the same effect). Generally speaking the major differences in functionality can be summarized as follows: SABLE offers finer-grained control for the specification of factors influencing prosody (e.g., several levels of strength in <emph>, several levels of size in <break>) as well as concrete control over the acoustic parameters (i.e., fundamental frequency in Hertz and duration in milliseconds) for specifying prosody.
Online References:
http://msdn.microsoft.com/library/default.asp?url=/library/en-us/sapi/Welcome.asp

SML in VHML

VHML (Virtual Human Markup Language) is an attempt to combine existing markup-languages developed for the various aspects of human-computer interaction (e.g. facial expression, body animation, emotional representation) into a unified specification language. The sub-part of VHML concerned with the markup for speech synthesis is called Speech Markup Language (SML) and is – according to the current "VHML Working Draft v0.3" from 21.Oct.2001 – based on W3C's SSML. Comparing SML to the current version of SSML points out, that roughly speaking SML currently is a slightly downsized variant of SSML. The most important difference to SSML and SAPI is, that it foresees the labelling of the speaker's emotion via VHML's Emotional Markup Language (EML). Emotion-tags specified in EML are inherited by SML and are thus visible to the speech synthesis.

Comparison of SSML, SML and SAPI

In Table 1 a rough comparison of functionality and tag-sets of SSML, its VHML-derivate SML, and MS-SAPI is performed. "N.A." denotes "Not available". Elements and their meaning are usually described in more detail in the section on SSML.

Functionality	**SSML**	**SML**	**SAPI**	**Remarks**
Root element	<speak>	N.A.	<sapi>	
Defining language	"lang" attribute for <speak> <sentence> and <paragraph>	N.A.	<lang>-element or "lang"-attribute of <voice> element	
Structuring of text	<paragraph> <sentence>	<paragraph>	N.A.	<vhml:paragraph> is not part of SML proper, but available in VHML
Specifying how a certain content is to be interpreted. Attributes are, e.g., *email, number, ordinal*	<say-as>	<say-as>	<context>	

Functionality	SSML	SML	SAPI	Remarks
Specify, that contained text is to be spelled out (e.g. “USA”)	<say-as type=spell-out>		<spell>	
Provide part of speech information	N.A.	NA	<PartOfSp>. Values are: “Unknown”, “Noun”, “Verb”, “Modifier”, “Function”, “Interjection”	
Provide phonetic pronunciation	<phoneme>	<phoneme>	<pron>	
Indicate that a specified text substitutes another	<sub>	N.A.	N.A.	
Specify/change the voice used	<voice> Attributes are lang, gender, age, variant, name	<voice>	<voice>	
Mark content as emphasized	<emphasis> Attribute: level [Values: “strong”, “moderate”, “none”, “reduced”]	<emphasize-syllable> extends SSML with Attribute: affect [Values “pitch”, “duration”, “both”] and target [the phoneme to be emphasized]	<emph>	In SAPI only <EMPH> available w/o further differentiation. In SML the syllable and optionally the prosodic means to signal emphasis and the exact position of the phoneme to be emphasized can be specified.
Control of pausing/prosodic boundaries.	<break> Attribute: size [Values: “none”, “small”, “medium”, “large”] Attribute: time (optional): duration of pause	<break>	<silence msec=100> Inserts silence of 100ms length	In SAPI only length of pause specified – no further differentiation
Control prosody.	<prosody> Attributes: pitch, contour, range, rate, duration, volume	<prosody>	<pitch> <rate><speed> <volume>	SSML allows for a finer grained control of prosody. Equivalents to “contour” and “duration” are missing entirely in

Functionality	SSML	SML	SAPI	Remarks
				SAPI
Insert arbitrary audio file (e.g., recorded speech, music)	<audio>	<embed>	NOT AVAILABLE	<vhml:embed> is not part of SML proper but of VHML – it allows for embedding of foreign filetypes in general
Specify emotion of speaker	N.A.		N.A.	
Place a marker in the text: An event will be issued by the synthesizer when reaching the mark	<mark>	<mark>	<bookmark>	<ssml:mark> can have content, <sapi:bookmark> has to be empty

Table 1 Summary of all attributes in SSML, SML, and SAPI

This table once again reveals the close similarities between SSML and SML. SAPI basically offers a sub-set of SSML's labels. Many of the remaining differences are only syntactic.