ESTADO ACTUAL DEL ESTÁNDAR SSLM PARA LA SÍNTESIS DEL HABLA DENTRO DEL "SPEECH INTERFACE FRAMEWORK" DESARROLLADO POR EL GRUPO DE TRABAJO "VOICE BROWSER" DEL W3C

D. Mario Rodríguez Boya
Universidad de Oviedo
mariorb@gmail.com

RESUMEN

La especificación SSML (Speech Synthesis Markup Language) es uno de los estándares propuestos por el W3C que hacen posible el acceso a la Web mediante la interacción del habla. SSML es un lenguaje de marcado para la síntesis de texto en voz basado en XML. Este lenguaje tiene como principal objetivo ofrecer a los autores de contenido sintetizable un camino estándar para controlar aspectos del habla como pueden ser la pronunciación, volumen, tono y velocidad, mejorando así la calidad del contenido sintetizado.

Palabras clave: SSML, Síntesis de Voz, Síntesis del Habla, Navegación por Voz.

1. INTRODUCCIÓN

SSML [SSML] es un estándar desarrollado por el grupo de trabajo *Voice Browser* y consiste en un lenguaje de etiquetado basado en XML que, junto con otras especificaciones como VoiceXML o SRGS (Speech Recognition Grammar Specification), forman el denominado *W3C Speech Interface Framework*.

El conjunto de especificaciones desarrolladas cubre las necesidades de diálogo mediante voz, reconocimiento del habla, síntesis de voz, etc. Algunos ejemplos de las posibles aplicaciones de estos estándares son: servicios para las empresas como la automatización de respuestas por teléfono, soporte técnico, rastreo de pedidos, información de despegue y llegada de vuelos; acceso a la información pública como el servicio meteorológico o del estado del tráfico; información personal como calendarios o direcciones de teléfono; y asistentes personales para la comunicación con otra gente enviando y recibiendo correos electrónicos de voz.

El papel principal de SSML es suministrar a los desarrolladores de contenidos sintetizables un método estándar para controlar aspectos del habla como la pronunciación, tono, volumen y velocidad tanto en la Web como en otras aplicaciones.

Una iniciativa ligada al establecimiento de un sistema estándar fue SABLE [SABLE], que intentaba integrar distintos lenguajes de marcas basados en XML para sintetizar voz en uno solo. El trabajo llevado a cabo por SABLE sirvió de punto de partida para la definición de los requisitos de un lenguaje de marcas para la síntesis de voz.

SSML se pueden generar tanto automáticamente, por ejemplo a través de XSLT, como explícitamente por un autor humano. Además, puede presentarse como un documento completamente escrito en SSML o como fragmentos embebidos en otros lenguajes.

El proceso de diseño y estandarización ha surgido a partir del documento Speech Synthesis Markup Requirements for Voice Markup Languages [REQS] que define los principales criterios de diseño (consistencia, interoperabilidad, generalidad, internacionalización, legibilidad, implementabilidad, etc.)

2. PROCESO DE SÍNTESIS DEL HABLA

Para el proceso de síntesis es necesario un procesador que transforme el texto en voz el cual soporte SSML. La creación de un documento de texto en formato SSML que sirva como entrada a un procesador de síntesis puede ser automática, llevada a cabo por un autor humano o una combinación de ambas.

El procesamiento del documento SSML ha de seguir una serie de pasos antes de llegar a generar la voz, son los siguientes:

1. Análisis gramatical del documento XML: utilizado para extraer el contenido del árbol. La estructura, etiquetas y atributos obtenidos en este paso influyen en cada uno de los pasos sucesivos.

2. Análisis de la estructura: la estructura de un documento afecta a la forma en la que el documento es leído. Las etiquetas <p> y <s> definidas explícitamente en SSML indican estructuras del documento que afectan directamente a la salida de voz, esto se denomina *Markup Support.* En los lugares del documento que estas marcas no son utilizadas es el procesador de síntesis el encargado de inferir la estructura mediante un análisis automatizado del texto, utilizando puntuación y demás elementos específicos del lenguaje, a esto se le denomina *Non-Markup Behaviour.*

3. Normalización del texto: todos los lenguajes tienen construcciones especiales que requieren una conversión especial de la forma escrita a la forma hablada (e.g. 1/2 o 100 €). Esta conversión es realizada automáticamente por el procesador de síntesis. El elemento `say-as` puede ser utilizado en el documento de entrada para indicar explícitamente la presencia de este tipo de construcciones. De esta forma se produce la desambiguación de términos como 1/2 que

puede tener múltiples significados (uno de dos, 1 de Enero, una mitad, etc.). Para el resto de texto que no esté marcado con `say-as` el procesador de síntesis es el encargado de realizar una conversión razonable. Debido a las ambigüedades es muy común que se produzcan errores en la transformación.

4. Conversión del texto a fonemas: una vez que el procesador de síntesis el conjunto de palabras que han de ser pronunciadas, se debe deducir la pronunciación de dichas palabras. La pronunciación de las palabras debe ser convenientemente descrita como secuencias de fonemas, que son unidades de sonido en un lenguaje que sirven para distinguir una palabra de otra. Cada lenguaje tiene conjunto específico de fonemas. Algunos lenguajes tienen entre 12 y 15 fonemas y otros más de 100. Hay lenguajes, como el inglés, en el que existe ambigüedad en la conversión del texto en voz (e.g. “read” y “reed” se pronuncian igual), la asociación no es uno a uno. En el caso del español esta conversión sí es uno a uno, por lo que este proceso es más sencillo. SSML proporciona el elemento `phoneme` para que el autor pueda controlar de forma explícita la pronunciación. En ausencia del elemento `phoneme` el analizador de síntesis aplicará normas para la pronunciación. Esto es llevado a cabo normalmente buscando las palabras en un diccionario de pronunciación que es dependiente del lenguaje.

5. Análisis prosódico: la prosodia es el conjunto de rasgos del habla que incluye el tono, el ritmo, las pausas, la velocidad y el énfasis. Conseguir una prosodia humana es importante para lograr una voz natural y comprensible. Para conseguir estos rasgos explícitamente SSML proporciona los elementos `emphasis`, `break` y `prosody`. En ausencia de estos elementos el procesador de síntesis establece estos rasgos de una forma bastante efectiva en la actualidad por defecto.

6. Generación de la forma de onda: los fonemas y la información prosódica son utilizados por el procesador de síntesis para producir la forma de onda correspondiente. SSML proporciona el elemento `voice` para solicitar una voz específica con unas determinadas cualidades (e.g. voz de una mujer joven).

SSML proporciona una forma estándar de especificar las propiedades de la voz sintetizada como la pronunciación, volumen, tono, etc. Estos valores son meras indicaciones para el procesador de síntesis por lo que la decisión final sobre sus valores la tiene este último en caso de que no le parezcan razonables.

3. FORMATO DEL DOCUMENTO

Un documento SSML ha de tener un encabezado XML [XML]. El DOCTYPE debe ser como sigue:

```
<!DOCTYPE speak PUBLIC "-//W3C//DTD SYNTHESIS 1.0//EN"
  "http://www.w3.org/TR/speech-synthesis/synthesis.dtd">
```

El encabezado es seguido por el elemento raíz `speak`. Éste elemento ha de designar el espacio de nombres mediante el atributo `xmlns`. También es recomendado que el elemento `speak` indique la localización del esquema SSML mediante el atributo `xsi:schemaLocation`. A continuación se presentan dos ejemplos de encabezados válidos en SSML.

Ejemplo 1:

```
<?xml version="1.0"?>
<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">
```

Ejemplo 2:

```
<?xml version="1.0"?>
<!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"
         xml:lang="en-US">
```

Los elementos `meta`, `metadata` y `lexicon` deben ir antes de cualquier otro elemento y del texto contenido dentro del elemento raíz `speak`. No existe ninguna otra restricción del orden de los elementos en la especificación.

Un uso típico de SSML es la lectura de los correos electrónicos. En el siguiente ejemplo se leen los encabezados de los mensajes. Los elementos `<p>` y `<s>` se usan para establecer la estructura del texto. El elemento `<break>` se sitúa antes de de la hora para dar énfasis en la importancia de esta información y que el oyente preste atención. El elemento `<prosody>` se utiliza para hacer que la velocidad del habla sea más lenta cuando se lee el asunto del mensaje y así el oyente pueda realizar anotaciones de los detalles con más tranquilidad.

```
<?xml version="1.0"?>
<!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">
  <p>
    <s> Tienes 4 mensajes.</s>
    <s> El primero es de Mario, recibido a las <break/>
3:45pm.</s>
    <s> El asunto es <prosody rate="-
20%">vacaciones</prosody></s>

  </p>
</speak>
```

También es posible en SSML combinar ficheros de audio y diferentes voces. En el siguiente ejemplo se suministra información de una colección de música turnando voces masculinas y femeninas.

```
<?xml version="1.0"?>
<!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. INTEGRACIÓN CON OTROS LENGUAJES DE MARCAS

SMIL (Synchronized Multimedia Integration Language) [SMIL] facilita la creación de presentaciones audiovisuales. SMIL es usado normalmente para presentaciones multimedia las cuales integran audio, video, texto o cualquier otro elemento multimedia. SSML y SMIL se complementan especialmente bien en la en la descripción de aplicaciones multimedia dinámicas que incluyen salida mediante síntesis de voz. Una muestra de lo anterior lo vemos con el siguiente ejemplo.

Tenemos el fichero 'greatings.ssml' siguiente escrito en SSML:

```
<?xml version="1.0"?>
<!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">

  <s>
    <mark name="greetings"/>
    <emphasis>Greetings</emphasis> from the <sub alias="World
Wide Web Consortium">W3C</sub>!
  </s>
</speak>
```

Y el siguiente fichero SMIL:

```
<smil xmlns="http://www.w3.org/2001/SMIL20/Language">
  <head>
    <top-layout width="640" height="320">
      <region id="whole" width="640" height="320"/>
    </top-layout>
  </head>
  <body>
    <par>
      <img src="http://www.w3.org/Icons/w3c_home"
region="whole" begin="0s"/>
      <ref src="greetings.ssml" begin="1s"/>
    </par>
  </body>
</smil>
```

Este último muestra el logotipo del W3C y un segundo después se genera el discurso establecido en el fichero 'greeting.ssml'.

ACSS (Aural Cascading Style Sheets) [CSS2] se ocupa de enriquecer las formas visuales de documentos (como xhtml) con elementos adicionales que ayudan en la síntesis de texto en audio. En comparación con SSML, los

documentos generados por ACSS tienen mayor capacidad para especificar propiedades de la secuencia de audio, incluyendo la designación de la localización en 3D de fuentes de audio. Muchos otros elementos ACSS superan la funcionalidad de SSML, especialmente en la especificación del tipo y calidad de la voz. SSML debe entenderse como un superconjunto de las capacidades de ACSS exceptuando las funcionalidades de audio espacial.

VoiceXML (Voice Extensible Markup Language) [VXML] permite el desarrollo basado en Web de aplicaciones de respuesta interactivas mediante voz (navegadores de voz). VoiceXML soporta la síntesis de voz, grabado y reproducción de audio digital, reconocimiento del habla, entrada DTMF (Dual Tone Multi Frecuency) y control de llamadas telefónicas. VoiceXML 2.0 extiende SSML para el marcado de texto para ser sintetizado. En el siguiente ejemplo se muestra la integración de ambos lenguajes.

```
<?xml version="1.0" encoding="UTF-8"?>
<vxml version="2.0" xmlns="http://www.w3.org/2001/vxml"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.w3.org/2001/vxml
   http://www.w3.org/TR/voicexml20/vxml.xsd">
   <form>
      <block>
         <prompt>
           <emphasis>Welcome</emphasis> to the Bird Seed
Emporium.
           <audio
src="rtsp://www.birdsounds.example.com/thrush.wav"/>
           We have 250 kilogram drums of thistle seed for
           $299.95
           plus shipping and handling this month.
           <audio
src="http://www.birdsounds.example.com/mourningdove.wav"/>
         </prompt>
      </block>
   </form>
</vxml>
```

5. IMPLEMENTACIONES

Existen algunas implementaciones disponibles de SSML, algunas de ellas de código abierto y otras propietarias.

Loquendo TTS soporta completamente la especificación SSML 1.0.

Microsoft Speech Server soporta SSML para la síntesis de texto en voz. Microsoft Speech Application SDK es un conjunto de herramientas para la implementación de aplicaciones de telefonía por voz y aplicaciones multimodal.

OptimTalk proporciona métodos para la integración con SSML incluso cuando los procesadores de síntesis no tienen soporte nativo SSML (e.g. Microsoft Speech API 5.1). El soporte SSML podría limitarse en cierta

medida por las habilidades de un procesador determinado en este caso. Algunos procesadores del habla han sido integrados.

Voxeo Corporation ofrece la plataforma VoiceCenter™ IVR que también soporta la especificación SSML 1.0.

Como implementación de código abierto cabe destacar el proyecto FreeTTS.

6. DESAFÍOS FUTUROS

Uno de los desafíos a los que se enfrenta el proceso de síntesis del habla es la normalización del texto debido a los múltiples homógrafos (palabras con múltiples significados), abreviaturas y símbolos utilizados en los lenguajes escritos y que deben ser transformados en los correspondientes fonemas.

Muchos sistemas de texto a voz no generan representaciones semánticas de los textos de entradas, pues los sistemas para hacerlo no son fiables o computacionalmente efectivos. Como resultado, se usan varias técnicas heurísticas para estimar la manera correcta de desambiguar homógrafos, como buscar palabras vecinas y usar estadísticas sobre la frecuencia de aparición de las palabras.

Otro gran desafío es la implantación generalizada de servicios y navegadores de voz en todos los ámbitos para facilitar la interacción entre computadoras y seres humanos; y que éstos sean desarrollados de manera estandarizada.

7. CONCLUSIONES

El trabajo del *W3C Voice Browser Working Group* ha dado lugar a un conjunto de estándares que ha hecho posible una interacción hombre-máquina más natural. Concretamente el trabajo realizado en la especificación SSML 1.0 ha dado lugar a la implementación de aplicaciones de síntesis del habla que generan voces casi totalmente humanas.

El problema al que se enfrenta un autor de documentos SSML actualmente es decidir si deja en manos del procesador de síntesis la mayor parte del proceso de síntesis o si por el contrario controla explícitamente todos los detalles.

En el futuro los sistemas deberán inferír sin lugar a error el significado de cada palabra dependiendo de su contexto y de esta manera interactuar de una manera más efectiva y natural con las personas.

7. REFERENCIAS

[SSML] Speech Synthesis Markup Language (SSML) Version 1.0, W3C Voice Browser Working Group. http://www.w3.org/TR/speech-synthesis/

[XML] *Extensible Markup Language (XML) 1.0 (Second Edition)*, T. Bray et al., Editors. World Wide Web Consortium, 6 October 2000. This version of the XML 1.0 Recommendation is http://www.w3.org/TR/2000/REC-xml-20001006. The latest version of XML 1.0 is available at http://www.w3.org/TR/REC-xml.

[SABLE] "SABLE: A Standard for TTS Markup", Richard Sproat, et al. *Proceedings of the International Conference on Spoken Language Processing*, R. Mannell and J. Robert-Ribes, Editors. Causal Productions Pty Ltd (Adelaide), 1998. Vol. 5, pp. 1719-1722. Conference proceedings are available from the publisher at http://www.causalproductions.com/.

[REQS] *Speech Synthesis Markup Requirements for Voice Markup Languages*, A. Hunt, Editor. World Wide Web Consortium, 23 December 1999. This document is a work in progress. This version of the Synthesis Requirements is http://www.w3.org/TR/1999/WD-voice-tts-reqs-19991223/. The latest version of the Synthesis Requirements is available at http://www.w3.org/TR/voice-tts-reqs/.

[SMIL] *Synchronized Multimedia Integration Language (SMIL 2.0)*, J. Ayars, et al., Editors. World Wide Web Consortium, 7 August 2001. This version of the SMIL 2 Recommendation is http://www.w3.org/TR/2001/REC-smil20-20010807/. The latest version of SMIL2 is available at http://www.w3.org/TR/smil20/.

[VXML] *Voice Extensible Markup Language (VoiceXML) Version 2.0*, S. McGlashan, et al., Editors. World Wide Web Consortium, 16 March 2004. This version of the VoiceXML 2.0 Recommendation is http://www.w3.org/TR/2004/REC-voicexml20-20040316/. The latest version of VoiceXML 2 is available at http://www.w3.org/TR/voicexml20/.

[CSS2] *Cascading Style Sheets, level 2: CSS2 Specification*, B. Bos, et al., Editors. World Wide Web Consortium, 12 May 1998. This version of the CSS2 Recommendation is http://www.w3.org/TR/1998/REC-CSS2-19980512/. The latest version of CSS2 is available at http://www.w3.org/TR/REC-CSS2/.

http://www.w3.org/TR/2004/REC-speech-synthesis-20040907/

http://www.naturalvoices.att.com/

http://public.research.att.com/~ttsweb/tts/demo.php

http://www.loquendo.com/en/technology/TTS.htm

http://www.microsoft.com/speech/

http://www.optimsys.cz/

http://www.voxeo.com/

http://es.wikipedia.org/wiki/Comunicaci%C3%B3n_multimodal

http://en.wikipedia.org/wiki/SSML

http://en.wikipedia.org/wiki/voiceXML

http://www.xml.com/pub/a/2004/10/20/ssml.html

http://msdn.microsoft.com/msdnmag/issues/06/01/speechinWindowsVista/