NATURTTSML – UN ESQUEMA DE ANOTACIÓN PARA LA MEJORA DE LA NATURALIDAD EN LOS SISTEMAS DE SÍNTESIS DE VOZ

Albert González Lamaña
Universidad Nacional de Educación a Distancia (UNED), Madrid
Antonio Pareja Lora
Universidad Complutense de Madrid / ATLAS (UNED), Madrid

RESUMEN

Aunque el resultado de los sistemas de conversión de texto a voz (CTV) actuales suele ser bastante inteligible, un CTV produce habitualmente una lectura monótona. Una de las principales causas de este problema es que un CTV no puede interpretar el contenido de un texto plano. Por lo tanto, para conseguir una lectura más expresiva y natural es necesario indicarle cómo debe leer el texto. Existen varios lenguajes y esquemas de anotación para este fin, pero parciales y/o centrados en diferentes aspectos del habla expresiva. Este artículo presenta un esquema de anotación (NaturTtSML) que reúne en uno solo los elementos principales de todos ellos.

Palabras clave: conversor de texto a voz, anotación, emociones, habla expresiva, narración de cuentos

ABSTRACT

Even though the output of current Text-to-Speech (TtS) systems is usually quite comprehensible, this output is quite often too monotonic. One of the main causes for this problem is that a TtS normally cannot actually understand plain text. Thus, the TtS system must be provided with the way to read it, in order to supply a more natural and expressive speech. So far, some annotation languages and schemes have been developed towards this end; however, they are partial and/or focus in different areas of expressive speech. This paper presents an annotation scheme (NaturTtSML) that merges and makes interoperate all of them altogether in just one scheme.

Keywords: TtS (Text-to-Speech), annotation, emotions, expressive speech, storytelling

1. INTRODUCCIÓN

Actualmente, la síntesis de voz ofrece voces de gran calidad e inteligibilidad. A pesar de ello, en muchas ocasiones la falta de variación y de expresividad hace que las voces suenen monótonas, como por ejemplo, cuando se realiza una lectura de un texto literario, como los cuentos infantiles usados en el presente trabajo.

Desde el desarrollo de los primeros conversores de texto a voz (CTV) se ha intentado mejorar la naturalidad y la expresividad de las voces. Numerosos estudios en este campo se han centrado en la expresión de emociones, utilizando diferentes técnicas y/o tecnologías: síntesis por formantes (Cahn, 1990; Murray y Arnott, 1995; Burkhardt, 2001), síntesis por concatenación de unidades (Bulut *et al.*, 2002, Esquerra *et al.*, 2004; Iida *et al.*, 2000; Iida y Campbell, 2003; Iriondo *et al.*, 2000; Johnson *et al.*, 2004) o síntesis basada en Modelos Ocultos de Markov (Yamagishi *et al.*, 2003).

Sin embargo, por un lado, la naturalidad en el habla no sólo está relacionada con la expresión de emociones, sino también con el uso de una entonación adecuada. Además del estado anímico, otros aspectos importantes son, por ejemplo, la modalidad oracional del enunciado o la intención comunicativa del hablante.

Por otro lado, uno de los problemas para obtener una lectura más natural y expresiva es la falta de información explícita en los propios textos. Un CTV no puede, por ejemplo, diferenciar cuándo habla un personaje u otro, cuál es su estado emocional o cuál es la intención real de su mensaje.

La solución a este problema pasa por anotar la información que cualquier ser humano puede interpretar de manera implícita de forma que pueda ser entendida por un CTV. Actualmente, existen diferentes lenguajes para anotar diferentes aspectos, como las emociones o los actos de habla. Sin embargo, no existe ninguno que los englobe a todos.

Por este motivo, partiendo de las directrices de la *Text Encoding Iniciative* (TEI), se ha creado un esquema de anotación (NaturTtSML) basado en las recomendaciones XML (Bray *et al.*, 2008) y XML *Schema* (Biron, Malhotra y *World Wide Web Consortium*, 2004) que toma elementos de esquemas y lenguajes de anotación existentes para la síntesis de voz (SSML − Taylor e Isard (1997)), para la anotación de emociones (EmotionML − Schröder *et al.* (2011)) y para la anotación pragmática (ISO 24617-2:2012 (*SemAF - Part 2: Dialogue acts*)).

2. DESCRIPCIÓN DEL ESQUEMA NATURTTSML

El esquema de anotación NaturTtSML parte de los módulos y elementos proporcionados por las directrices de la TEI. La estructura básica de un documento TEI está compuesta por tres elementos: `<TEI>`, `<teiHeader>` y `<text>`.

El elemento `<text>` contiene el documento que se va a anotar y puede contener tres elementos: `<body>` (texto del documento), `<front>` (elementos que preceden al texto, como por ejemplo un prólogo) y `<back>` (elementos que aparecen después del texto, como apéndices o índices).

Además de los cuatro módulos básicos recomendados por las directrices de la TEI (módulo **`tei`**, módulo **`core`**, módulo **`header`**, módulo **`textstructure`**), en el esquema se hace uso de los módulos **`drama`**, **`spoken`** y **`linking`**, cuyos elementos y usos se describen en los siguientes apartados.

```
<TEI>
  <teiHeader>
    <!--...-->
  </teiHeader>
  <text>
    <!--...-->
  </text>
</TEI>
```

Figura 1. Estructura básica de un documento TEI

2.1. Módulo **`drama`**

Este módulo está destinado a la codificación y anotación de textos teatrales. Del mismo modo que en los cuentos, en este tipo de textos intervienen uno o más personajes. En nuestro caso, los textos van a ser "interpretados" por las voces de un CTV, que serán nuestros "actores".

El elemento `<castList>` del módulo **`drama`**, situado en el elemento `<front>`, se usa para presentar a los personajes de una obra. La información de cada uno de ellos está contenida dentro de un elemento `<castItem>`, especi-

ficada mediante tres elementos: <role>, <roleDesc> y <actor>. Un ejemplo de la estructura básica del elemento <castList> y de los elementos que contiene se puede ver en la Figura 2.

2.1.1. El elemento <role>

```
<castList>
   <castItem>
      <role/>
      <roleDesc/>
      <actor/>
   <castItem/>
<castList/>
```

Figura 2. Estructura básica de un elemento <castList>

Contiene el nombre del personaje o el rol que tiene en la historia. Asimismo, posee el atributo xml:id, con el que se le otorga un identificador a cada personaje, que será referenciado en cada una de sus intervenciones.

2.1.2. El elemento <roleDesc>

El elemento <roleDesc> sirve para describir las características de los personajes. Los atributos que admite este elemento son: sex (sexo), age (edad), kind (tipo de personaje) y personality (carácter) − véase la Tabla 1.

Atributos	Valores
sex	**male**, **female**, **neutral**
age	**child**, **young**, **adult**, **senior**
kind	**human**, **animal**, **creature**, **naturalObject**, **artefact**
personality	**good**, **bad**, **neutral**

Tabla 1. Atributos y valores del elemento <roleDesc>

2.1.3. El elemento `<actor>`

Mediante este elemento es posible individualizar a los personajes, por ejemplo, mediante la asignación de una voz femenina o masculina, una velocidad de habla específica o un tono determinado. La lista completa de atributos se puede ver en la Tabla 2.

Atributos del elemento **<actor>**
tool
voiceName
pitch
pitchRange
speechRate
pauseDuration
intensity
voiceQuality

Tabla 2. Lista de atributos del elemento <actor>

Los dos primeros atributos, `tool` y `voiceName`, hacen referencia, por un lado, al CTV que se va a usar y, por otro, al nombre de la voz utilizada en dicha herramienta. Los otros atributos, de carácter optativo, se usarán para modificar los valores por defecto de la voz.

Debido a que ya existe un lenguaje estándar (SSML) que permite la anotación de muchos de estos parámetros, los valores admitidos para los diferentes atributos siguen las especificaciones de dicho estándar. El número de atributos incluidos en el esquema que se quieran especificar para el personaje dependerá del CTV utilizado y de la manera en que puedan ser modificados los parámetros.

En la Figura 3 se puede observar un ejemplo de anotación de un elemento `castItem`.

```
<castItem>
        <role xml:id=”NAR”>narrador</role>
        <roleDesc sex=”male”
                age=”adult”
                kind=”human”
                personality=”neutral”/>
        <actor tool=”Festival TTS”
                voiceName=”JuntaDeAndalucia_es_pa_diphone”
                pitchRange=”10%”
                speechRate=”-60%”
                PauseDuration=”275%”/>
</castItem>
```

Figura 3. Ejemplo de anotación de un elemento `castItem`

2.2. El módulo **spoken**

El módulo **spoken** contiene la descripción de etiquetas adecuadas para la anotación de transcripciones orales. Los dos elementos de este módulo que se han utilizado son el elemento `<u>` (*utterance*) y el elemento `<pause>`.

2.2.1. El elemento `<u>`

Es el elemento principal para la organización del texto en el módulo **spoken** de la TEI; se define como un fragmento de habla precedido y seguido por una pausa, o por un cambio de hablante.

El elemento `<u>` posee dos atributos: `xml:id` (proporciona un identificador para la intervención) y `who`, que permite relacionar el elemento `<u>` con uno de los personajes del cuento, asignándole como valor una referencia al valor del atributo `xml:id` del elemento `<castItem>`. En la Figura 4 se puede ver un ejemplo de anotación de una intervención.

```
<u xml:id="utt6" who="#NAR">
      <seg   xml:id="seg61"
             emotion="neutral"
             sentenceType="declarative">Era éste el juego
que más la divertía.
      </seg>
</u>
```

Figura 4. Ejemplo de anotación de una intervención

2.2.2. El elemento `<pause>`

El otro elemento que se ha tomado del módulo **spoken** es `<pause>`, usado para insertar pausas entre dos elementos `<u>`.

2.3. El módulo **linking**

El módulo **linking** comprende una serie de recursos con los que se pueden representar diferentes análisis de la estructura de un texto.

2.3.1. El elemento `<seg>`

El elemento `<seg>` se usa para (a) dividir la intervención de un personaje en segmentos más pequeños; y (b) poder indicar, mediante una serie de atributos, cambios en la entonación:

a) Atributo `sentenceType`: permite anotar la modalidad oracional de la intervención.

b) Atributo `dialogueAct`: permite anotar el acto de habla o función comunicativa que está presente en la intervención que se quiere anotar.

c) Atributo `emotion`: permite anotar el tipo de emoción con la que debe ser leída la intervención.

4. SÍNTESIS DE LOS CUENTOS

El esquema NaturTtSML se utilizó para anotar un pequeño corpus de cuentos infantiles. La herramienta elegida para la conversión del texto a voz fue Festival. Los motivos principales para su elección fueron su gratuidad, el hecho de disponer de dos voces en español (masculina y femenina) y ser un programa de código abierto (*open source*).

La unión entre las anotaciones del cuento y la síntesis de voz de Festival se logró mediante el uso de unos scripts escritos en el lenguaje de programación Python, que permitieron transformar la información contenida en las anotaciones para que pudiera ser interpretada por Festival.

Las modificaciones efectuadas en la síntesis original tenían el objetivo de crear un estilo propio para cada personaje, adaptando la voz a su edad, sexo y rol en la historia. Además, se añadió a la entonación de Festival, declarativa por defecto, una entonación interrogativa (en los casos en los que era necesario). Finalmente, se simularon diferentes emociones modificando varios parámetros acústicos.

5. EVALUACIÓN Y RESULTADOS

5.1. Cuestionario

Con el objetivo de evaluar los audios de los cuentos sintetizados se creó un cuestionario con la aplicación *Google Forms*, accesible vía web en https://goo.gl/forms/OIull0HoI9W5xDbg1.

Las 9 preguntas que formaban la encuesta se dividieron en tres secciones:

1. Evaluación de personajes.
2. Evaluación de las entonaciones interrogativas.
3. Evaluación de la expresión de emociones.

Todas las preguntas estaban compuestas por pruebas de percepción, con tareas de tres tipos: identificación, comparación y valoración. En cada pregunta, los evaluadores debían escuchar uno o dos audios extraídos del corpus de cuentos, algunos en la versión original de Festival y otros en la versión modificada.

Un total de 17 evaluadores voluntarios y anónimos (4 hombres y 13 mujeres), con edades comprendidas entre los 25 y los 44 años, y en general con estudios de Filología (64,1%), tomaron parte en la evaluación. La mayoría manifestó no haber utilizado nunca un CTV (65%), un 29% lo había utilizado alguna vez y sólo uno de los encuestados (6%) respondió que utilizaba este tipo de tecnología frecuentemente.

5.2. Resultados

Los resultados obtenidos en la evaluación de las voces de los personajes han sido aceptables, a pesar de la baja calidad de las voces disponibles en Festival para el español y de las distorsiones producidas por la modificación de parámetros como el tono o el rango tonal. Los evaluadores fueron capaces de identificar el sexo y la edad de los personajes (Figura 5 y Figura 6) o de diferenciar a los personajes por su voz (Figura 7).

Por otro lado, los evaluadores consideraron que las voces modificadas eran más adecuadas para la lectura de cuentos infantiles que las originales de Festival. Esta evaluación se basaba en comparar la velocidad de habla, la duración y el número de pausas de un audio original de Festival y uno modificado (Figura 8, Figura 9 y Figura 10).

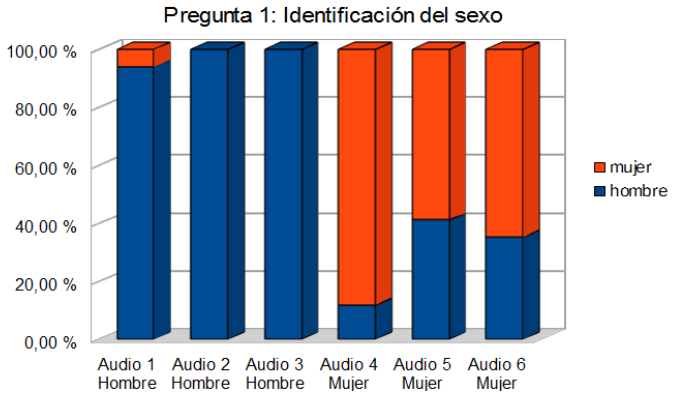

Figura 5. Identificación del sexo

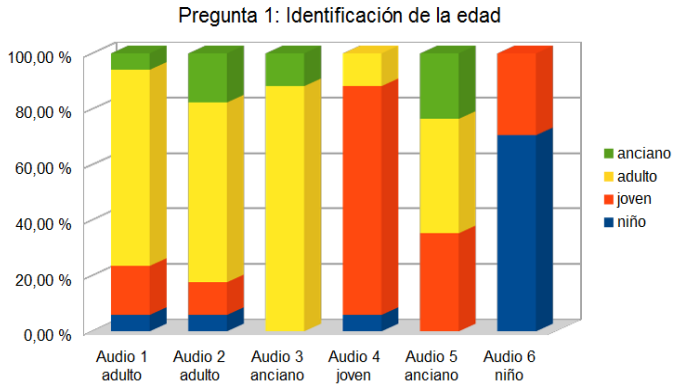

Figura 6. Identificación de la edad

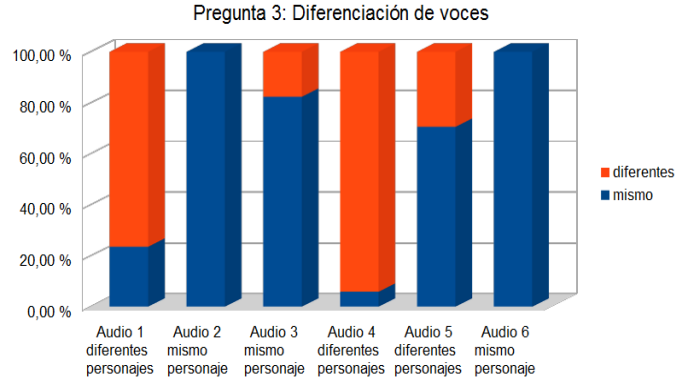

Figura 7. Diferenciación de personajes

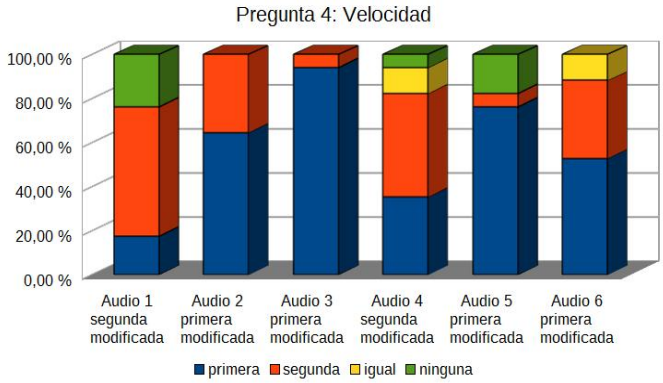

Figura 8. Comparación de la velocidad de habla

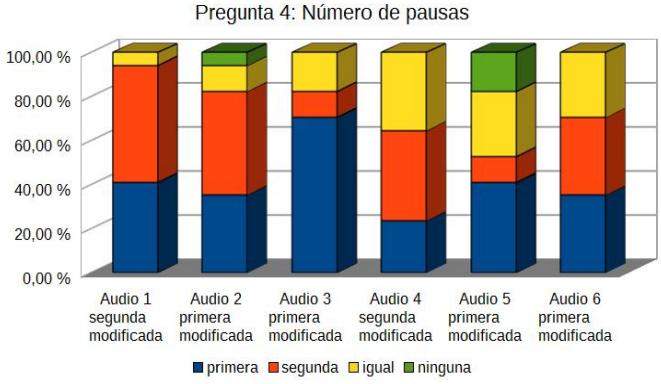

Figura 9. Comparación del número de pausas

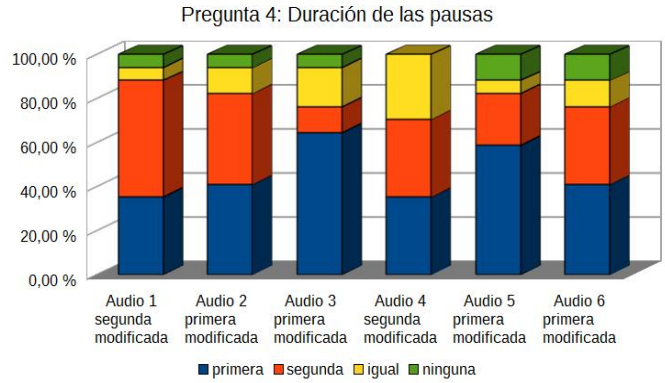

Figura 10. Comparación de la duración de las pausas

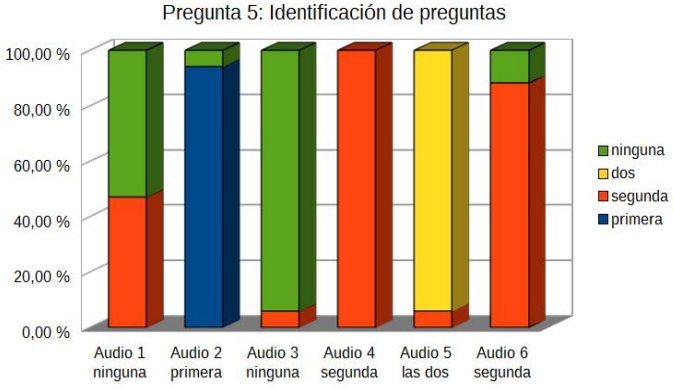

Figura 11. Identificación de enunciados interrogativos

En cuanto a la generación de entonaciones interrogativas, los evaluadores fueron capaces de diferenciar entre enunciados declarativos e interrogativos (Figura 11). Además, la valoración de la naturalidad de las interrogaciones fue considerada como aceptable.

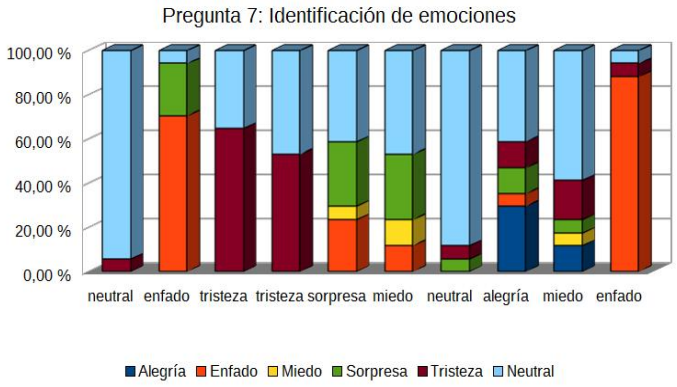

Figura 12. Identificación de emociones

Finalmente, la parte dedicada a la expresión de emociones fue la que obtuvo perores resultados menos positivos. La identificación de las emociones no ha fue satisfactoria, con buenos resultados solamente para el enfado y la tristeza, y resultados negativos para la sorpresa, la alegría o el miedo (Figura 12). La imposibilidad de poder modificar parámetros como la intensidad o la cualidad de voz en el CTV, junto con la baja calidad de las voces, pueden ser la explicación de los resultados obtenidos.

6. CONCLUSIONES Y TRABAJOS FUTUROS

Como se ha visto, para lograr una lectura más natural y expresiva es necesario que el CTV comprenda el contenido del texto. Por este motivo es necesario anotar aquellos elementos necesarios para obtener una entonación adecuada. El esquema de anotación NaturTtSML presentado cumple esta función. En él se han fusionado varios lenguajes y esquemas que han sido de utilidad para describir a los personajes que intervienen en un cuento y las características de sus voces. Asimismo, ha permitido segmentar el texto en enunciados y relacio-

narlos con cada uno de los personajes, y anotar emociones, modos oracionales y actos de habla.

Partiendo de los textos anotados y realizando las modificaciones oportunas en el sintetizador Festival para que pudiera interpretarlos, se ha obtenido una lectura dramatizada de los cuentos con voces propias para cada personaje. Además, se ha logrado modificar la entonación para realizar enunciados interrogativos, expresar correctamente algunas emociones y conseguir un estilo bastante adecuado para la lectura de cuentos infantiles.

Como trabajos futuros en relación al esquema de anotación, sería interesante diseñar una guía de anotación para que pudiera ser utilizado por varios anotadores y ver el grado de acuerdo entre ellos. Además, es necesaria la automatización parcial o total de la anotación, ya que hacerlo manualmente supone una gran inversión de tiempo y esfuerzo. Finalmente, sería interesante crear un corpus más amplio de textos anotados, que fuera abierto y reutilizable.

En la parte de la síntesis de voz, teniendo en cuenta las valoraciones obtenidas en la evaluación, sería importante reproducir las investigaciones utilizando una mayor variedad de voces y de mayor calidad. Asimismo, se debería mejorar la expresión de emociones, ya que este aspecto fue el peor valorado por los evaluadores.

7. AGRADECIMIENTOS

El presente trabajo se ha realizado en el contexto y/o con la ayuda de los proyectos **SWITCHED-ON** (nº ref.: FFI2016-80613-P), financiado por el Ministerio de Economía y Competitividad, y **eLITE-CM** (Edición LITeraria Electrónica, nº ref.: H2015/HUM-3426), financiado por (a) el Programa de Actividades de I+D entre grupos de investigación de la Comunidad de Madrid en Ciencias Sociales y Humanidades; y (b) *The European Social Funds* (*European Commission*) durante el periodo 2014-2020.

8. REFERENCIAS

Biron, P. y Malhotra, A. (2004). XML schema part 2: Datatypes. *World Wide Web Consortium Recommendation REC-xmlschema-2-20041028*.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., y Yergeau, F. (1998). Extensible markup language (XML), *World Wide Web Consortium*

Recommendation REC-xml-19980210, http://www. w3. org/TR/1998/REC-xml-19980210, *16*, 16.

Bulut, M., Narayanan, S. S., & Syrdal, A. K. (2002). Expressive speech synthesis using a concatenative synthesizer. En *INTERSPEECH*.

Burkhardt, F. (2001). Simulation emotionaler sprechweise mit sprachsyntheseverfahren. Reihe Berichte aus der Kommunikationstechnik. Shaker, Aachen.

Cahn, J. E. (1990). *The generation of affect in synthesized speech. Journal of the American Voice I/O Society*, 8, pp. 1-19.

Esquerra, I. y Bonafonte, A. (2004). Habla emocional mediante metodos de re-sintesis y seleccion de unidades, *URSI 2004. Actas del XIX simposium nacional de la unión científica internacional de radio.* Universitat Ramon Llull, Barcelona.

Iida, A., Campbell, N. (2003). Speech database design for a concatenative text-tospeech synthesis system for individuals with communication disorders. *International Journal of Speech Technology*, 6, pp. 379–392.

Iida, A., Campbell, N., Iga, S., Higuchi, F., y Yasumura, M. (2000). A Speech Synthesis System for Assisting Communication, *ISCA Workshop on Speech & Emotion*, Northern Ireland, pp. 167-172.

International Organization for Standardization (ISO). (2012). Language resource management - Semantic annotation framework (SemAF) - Part 2: Dialogue acts [ISO 24617-2:2012].

Iriondo, I., Guaus, R., Rodriguez, A., Lazaro, P., Montoya, N., Blanco, J., Bernadas, D., Oliver, J., Tena, D. y Longhi, L. (2000). Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. En *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, Northern Ireland, UK, pp. 161-166.

Johnson, W.L., Narayanan, S.S., Whitney, R., Das, R., Bulut, M. y LaBore, C. (2002). Limited domain synthesis of expressivemilitary speech for animated characters. En P*roceedings of the 7th International Conference on Spoken Language Processing*, Denver, Colorado, USA.

Murray, I. R., y Arnott, J. L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16(4), pp. 369-390.

Schröder, M., Baggia, P., Burkhardt, F., Pelachaud, C., Peter, C. y Zovato, E. (2011). EmotionML–an upcoming standard for representing emotions and related states, en *Proceedings of the 4th international conference on Affective computing and intelligent interaction*, Springer, pp. 316-325.

Taylor, P. e Isard, A. (1997). SSML: A speech synthesis markup language. *Speech Communication*, *21*(1-2)

Taylor, P., Black, A., y Caley, R. (1998). The Architecture of the Festival Speech Synthesis System, en *3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, pp. 147-151.

TEI Consortium, eds. (2016). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.1.0. Last updated on 15th December 2016. TEI Consortium.

Yamagishi, J., Onishi, K., Masuko, T., & Kobayashi, T. (2003). Modeling of various speaking styles and emotions for HMM-based speech synthesis. En *Proc. the 8th Eurpean Conference on Speech Communication and Technology* (No. III, pp. 2461-2464).