GENERACIÓN DE UNA VOZ SINTÉTICA EN CASTELLANO BASADA EN HSMM PARA LA EVALUACIÓN ALBAYZÍN 2008: CONVERSIÓN TEXTO A VOZ

R. Barra-Chicote[1]*, J. Yamagishi*[2]*, J. M. Montero*[1]*, S. King*[2]*, S. Lufti* [1]*, J. Macias-Guarasa*[3]

Grupo de Tecnología del Habla, Universidad Politécnica de Madrid[1],
Center for Speech Technology Research, University of Edinburgh[2],
Universidad de Alcalá[3]

RESUMEN

Este artículo describe el proceso de generación de una voz en castellano utilizando el corpus *UPC ESMA* de UPC proporcionado por la *Evaluación Albayzín 2008: Conversión Texto a Voz*. Se ha implementado una voz basada en selección de unidades mediante el paquete *Multisyn* de *Festival* y otra basada en *Hidden Semi-Markov Models* (HSMM) mediante *HTS*. Tras una breve evaluación de la calidad de ambas voces, se detallan las características principales de la voz basada en HSMM, sistema final presentado a la evaluación.

1. INTRODUCCIÓN

La *Evaluación Albayzín 2008: conversión texto a voz* tiene como objetivo la evaluación de las técnicas de síntesis actuales aplicadas al castellano, del mimo modo que la competición Blizzard Challenge para inglés y chino mandarín.

Cada equipo participante debe proporcionar una voz generada a partir del corpus proporcionado en un plazo de 7 semanas. Posteriormente deben sintetizar un conjunto de ejemplos de test, que serán evaluados perceptualmente, de forma conjunta con los del resto de equipos, en términos de similaridad con la voz original, naturalidad e inteligibilidad.

2. CORPUS

El corpus UPC ESMA [1] proporcionado para la evaluación del sistema consiste en las grabaciones de un conjunto de textos leídos con estilo neutro por parte de una locutora profesional.

El corpus proporciona 506 frases fonéticamente balanceadas (30 minutos), 208 párrafos de longitud media fonéticamente balanceados (30 minutos) y 62 párrafos literarios de mayor longitud (45 minutos).

Además del audio, señal de voz y señal del laringógrafo, se cuenta con el texto de referencia, la trascripción fonética y un diccionario con la información léxica. Con el corpus se proporciona la segmentación fonética y la marcación automática de *pitch*. Adicionalmente se dispone de la marcación manual de un subconjunto de la base de datos.

Este trabajo ha sido parcialmente financiado por el M.E.C. y los proyectos proyecto ROBONAUTA (DPI2007-66846-C02-02), EDECAN (TIN2005-08660-C04-04).

3. ANÁLISIS LINGÜÍSTICO

Para la realización del análisis lingüístico se han utilizado las herramientas proporcionadas por *Festival* [2]. Se ha prescindido de la información proporcionada con la base de datos y se ha empleado un alfabeto propio, un silabicador y un conversión grafema-alófono incorporados a *Festival*. El alfabeto utilizado consta de 30 alófonos típicos en castellano, entre los que se incluye el silencio.

Los módulos incorporados a *Festival* para llevar a cabo el análisis lingüístico son:

- Módulo de preproceso y normalización, que trata la pronunciación de nombres propios, acrónimos, números romanos y cifras.
- Módulo conversor grafema-alófono, que a partir de reglas fonéticas extrae la secuencia de alófonos del texto.
- Módulo silabicador, que a partir de la trascripción fonética y basándose en reglas, estima automáticamente la división en sílabas.
- Módulo acentuador, que determina, a partir de reglas, las sílabas tónicas y átonas de la secuencia alofónica.
- Módulo categorizador, que únicamente diferencia del resto el conjunto de palabras función.

A partir del análisis lingüístico se han extraído un conjunto de 65 características lingüísticas. Algunas de las más relevantes son:

- **A nivel de alófono:** Alófono anterior al predecesor, predecesor, actual, posterior, siguiente al posterior, y la posición del alófono actual en la sílaba.
- **A nivel de sílaba:** nº de fonemas y acentuación de la sílaba anterior, actual y posterior; posición de la

sílaba dentro la palabra y del grupo fónico; y la vocal de la sílaba.

- **A nivel de palabra:** la categoría gramatical (*POS*) de la palabra anterior, actual y posterior; nº de sílabas de la palabra anterior, actual y posterior; posición dentro del grupo fónico desde el comienzo y desde el final; y la posición del grupo fónico dentro de la frase.
- **A nivel de grupo fónico:** Nº de sílabas y de palabras del grupo fónico anterior, actual y posterior, y tipo de entonación final.
- **A nivel de frase:** Nº de sílabas, de palabras y de grupos fónicos.

4. SELECCIÓN DE UNIDADES *VERSUS* SÍNTESIS HSMM

En este trabajo se ha implementado una voz basada en selección de unidades y otra basada en Semi-Modelos Ocultos de Markov (HSMM: *Hidden Semi Markov Models*); con el fin de evaluar la bondad de cada técnica aplicada al corpus de la evaluación. Ambas voces han utilizado como módulo de preproceso el explicado en el apartado anterior.

En el caso de la voz basada en selección de unidades se ha utilizado el motor *multisyn*[3] de *Festival*. Durante la generación de esta voz se han encontrado un conjunto de problemas que han dado lugar a las siguientes limitaciones:

- Se ha tenido que prescindir de los párrafos literarios en el entrenamiento de HMM para la segmentación automática del corpus, usando únicamente las frases y los párrafos fonéticamente balanceados.
- A pesar de normalizar la intensidad de los ficheros de audio, se comprobaron variaciones de intensidad en los ejemplos sintetizados.
- Dado el tamaño del corpus, no se dispone de la suficiente cobertura de contextos lingüísticos como para modelar de forma implícita la parte prosódica [3], afectando a la naturalidad de la voz.

La voz basada en HSMM ha sido generada mediante *HTS 2.1* [4]. Algunos de los aspectos que diferencian esta voz de la anterior y que a priori mejoran la calidad de la voz (a falta de una evaluación exhaustiva) son:

- La segmentación fonética es un proceso implícito en el entrenamiento de los HSMM. A diferencia de la segmentación con *multisyn*, en este caso se utiliza información referente a la fuente de excitación (log F0 y componente aperiódica), un mayor número de coeficientes cepstrales y mayor número de estados.

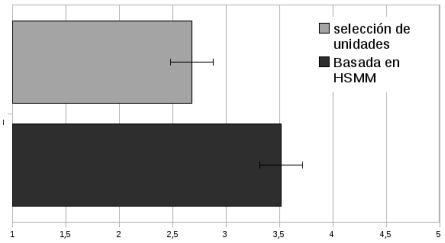

Figura 1. *Evaluación de la calidad de la voz basada en selección de unidades y la basada en HSMM.*

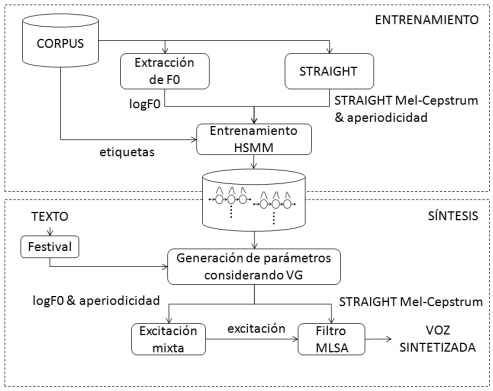

Figura 2. *Descripción del sistema (adaptada de [5]).*

- El uso de un modelo paramétrico proporciona mayor robustez, evitando discontinuidades. A priori, esta técnica proporciona una voz más estable y una síntesis más robusta para este volumen de datos de entrenamiento.

Se ha realizado una breve evaluación de calidad de las voces con objeto de seleccionar la mejor de ambas para la evaluación. 5 oyentes han evaluado 10 textos seleccionados del conjunto de ejemplos de test enviados por la organización de la evaluación, puntuando cada ejemplo siguiendo la escala MOS. Los resultados mostrados en la gráfica 1, indican que la calidad de la voz basada en HSMM (3,52) es mejor que la basada en selección de unidades (2,68).

5. CONVERSIÓN DE TEXTO A VOZ BASADA EN HSMM

En esta sección se describen las características principales del sistema empleado finalmente. Cada uno de los algoritmos empleados se detalla exhaustivamente en [5] y [4]. La Figura 2 presenta un diagrama general del sistema.

5.1. Modelo de producción de voz

Uno de los modelos de producción más extendidos es el conocido como *vocoder*. Este modelo consiste en modelar la voz humana como la convolución de un señal de

excitación con un filtro, el cual representa la información asociada al tracto vocal.

El uso de este modelo limita la calidad de la voz sintetizada, debido a que asume independencia entre la excitación y el filtro dado que simplifica la señal de excitación a un tren de impulsos en los sonidos sonoros, y a ruido en caso de los sonidos sordos. El resultado suele ser la percepción de una voz robótica.

Como solución a este problema, el sistema presentado incorpora STRAIGHT [6], vocoder que mejora la calidad de la síntesis al aplicar un procedimiento adaptativo sobre F0 en la estimación de la envolvente espectral. De esta forma se consigue separar la envolvente espectral de la componente periódica de la señal. Adicionalmente, se estiman medidas de aperiodicidad del espectro, basadas en la relación entre la zona de alta y de baja frecuencia de la envolvente espectral, las cuales representan la distribución relativa de energía de cada componente aperiódica [7].

En el proceso de síntesis, se utiliza un modelo de excitación mixta, basado en la suma de un tren de impulsos con manipulación de la fase y un ruido gausiano. La ponderación de ambas señales se realiza en el dominio de la frecuencia mediante las medidas de aperiodicidad comentadas anteriormente.

5.2. Entrenamiento de los modelos acústicos

Se ha utilizado una frecuencia de muestreo de $16kHz$ y un análisis trama a trama con un enventanado de tipo Blackman de $25ms$ y un desplazamiento de ventana de $5ms$.

Como ya se ha mencionado, en el sistema se han utilizado HSMMs para modelar la envolvente espectral, la información de aperiodicidad y el contorno de F0 (logaritmo de F0 realmente). Con el fin de que los modelos sean entrenables, es necesario codificar la información para disminuir la dimensionalidad de las observaciones. Para ello, a partir de la envolvente espectral se estiman los 40 primeros coeficientes cepstrales (*global mel cepstrum*) y las medidas de aperiodicidad se promedian en 5 subbandas de frecuencia.

Se ha prescindido de la información de las marcas de *pitch* proporcionadas con la base de datos. En nuestro sistema se ha buscado robustecer la estimación del contorno de logaritmo de F0 mediante el empleo de tres tipos de algoritmos de extracción de F0 a partir de la señal de voz. Finalmente, el contorno resultante es el promedio del resultado ofrecido por cada uno de los algoritmos por separado.

Adicionalmente, se calculan la primera y segunda derivada de cada una de las componentes estáticas, formando así un vector de 138 componentes.

En el caso de $logF0$ y sus derivadas se han modelado utilizado distribuciones MSD (*Multi Space Ditribution*) [8], en las que las tramas sonoras se modelan mediante un distribución gausiana con una matriz de covarianzas diagonal, y las tramas sordas mediante una distribución discreta.

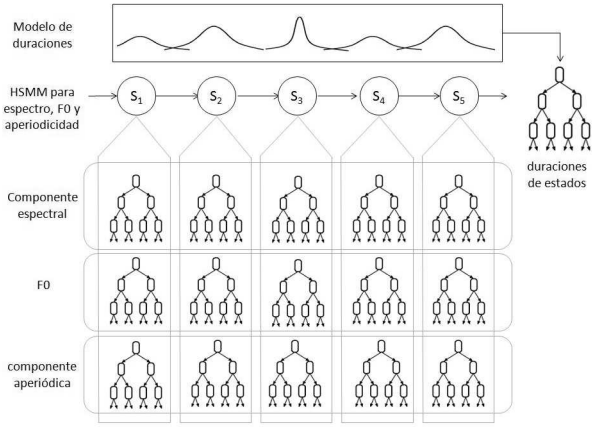

Figura 3. *HSMM dependientes del contexto (adaptada de [11]).*

5.2.1. Empleo de HSMM y modelado de duraciones

Los HSMM modelan la duración de cada estado de forma explícita mediante una función de distribución en lugar de utilizar las probabilidades de transición de los HMM convencionales, lo cual permite modelar el ritmo de una forma más apropiada [9].

En este caso se ha utilizado una función de distribución gausiana multivariable de dimensión equivalente al número de estados (5 en nuestro caso).

5.2.2. Modelos dependientes del contexto

Cada fonema se modela como un HSMM de 5 estados de izquierda a derecha. Para cada estado y cada una de las componentes del modelo (espectro, F0, aperiodicidad y duraciones) se entrenan, de forma independiente pero síncrona [10], un conjunto de modelos dependientes del contexto para cada estado. Éstos se estiman mediante el entrenamiento un árbol de decisión para cada componente aplicando un criterio basado en la *Minimum Description Length* (MDL).

En la generación del árbol de decisión, se ha partido de un conjunto inicial de 2042 preguntas relacionadas con el contexto a nivel fonético (se han utilizado pentafonemas), de sílaba, de palabra o grupo fónico.

El resultado es un conjunto de 63773 modelos para la componente espectral, logF0 y aperiódica y 17556 para el modelado de duraciones. La Figura 3 muestra el conjunto de modelos entrenados.

5.3. Generación de parámetros considerando su varianza global

La generación de secuencias de parámetros se lleva a cabo mediante el algoritmo introducido en [12]. Mediante

la relación entre las características estáticas y dinámicas se generan trayectorias suavizadas de parámetros.

Habitualmente, este suavizado suele ser excesivo, y para evitar esto se incorpora la varianza global de las características como parámetro de optimización junto al de la probabilidad de la observación dada la secuencia de parámetros. En [13] se describe en detalle la consideración de la varianza global en la generación de trayectorias.

5.4. Síntesis de voz

A la hora de sintetizar la señal de voz es necesario estimar la envolvente espectral. Dicha envolvente se aproxima mediante un filtro MLSA (*Mel Log Spectrum Approximation*), con el fin de reducir el coste computacional, estimado a partir de los coeficientes mel-cepstrum. La síntesis se realiza periodo a periodo como la convolución de una fuente de excitación mixta y dicho filtro MLSA [5].

6. CONCLUSIONES

Este trabajo describe la implementación de una voz sintética en castellano basada en HSMM para la *Evaluación Albayzín 2008: Conversión Texto a Voz*. Se han implementado voces basadas en las dos técnicas actuales que compiten en síntesis de voz, selección de unidades y síntesis basada en HSMM. Dichas voces se han implementado usando *Multisyn* de *Festival* y *HTS 2.1* respectivamente. Se ha realizado una evaluación limitada para decidir el mejor sistema para la competición, y finalmente se han descrito las características principales de cada uno de sus módulos. Una demostración de ambos sistemas se puede encontrar on-line en [14].

7. AGRADECIMIENTOS

Los autores agradecen a los miembros de CSTR y GTH su colaboración en la preparación de este trabajo.

8. BIBLIOGRAFÍA

[1] Antonio Bonafonte y Asuncion Moreno, "Documentation of the upc_esma spanish database," *TALP Research Center, Universitat Politecnica de Catalunya, Barcelona*, pp. 2781–2784, 2008.

[2] Paul Taylor, Alan W Black, y Richard Caley, "The architecture of the festival speech synthesis system," in *In The Third ESCA Workshop in Speech Synthesis*, 1998, pp. 147–151.

[3] Robert A. J. Clark, Korin Richmond, y Simon King, "Multisyn: Open-domain unit selection for the festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[4] The HTS working group, "Hmm-based speech synthesis system (hts). http://hts.sp.nitech.ac.jp," Último acceso: septiembre de 2008.

[5] H. Zen, T. Toda, M.Ñakamura, y K. Tokuda, "Details of nitech hmm-based speech synthesis system for the blizzard challenge 2005," *IEICE Trans. Inf. & Syst*, vol. E90-D, no. 1, pp. 325–333, January 2007.

[6] Hideki Kawahara, Haruhiro Katayose, Alain de Cheveign´e, y Roy D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity," *In Proc. of Eurospeech*, pp. 2781–2784, 1999.

[7] Hideki Kawahara, Jo Still, y Osama Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and syntesis system straight," *Proc MAVEBA*, pp. 13–15, September 2001.

[8] K. Tokuda, T. Masuko, N. Miyazaki, y T. Kobayashi, "Multi-space probability distribution hmm," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, March 2002.

[9] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, y T. Kitamura, "Hidden semi-markov model based speech synthesis," *In Proc. of ICSLP*, vol. II, pp. 1397–1400, October 2004.

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, y T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," *In Proc. of Eurospeech*, pp. 2347–2350, September 1999.

[11] Keiichi Tokuda, Heiga Zen, y Alan W. Black, "An hmm-based speech synthesis system applied to english," *Proc. of IEEE SSW*, vol. E90-D, no. 5, pp. 806–824, September 2002.

[12] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, y T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," *In Proc. of ICASSP*, pp. 1315–1318, June 2000.

[13] T. Toda y K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *IEICE Transactions*, vol. E90-D, no. 5, pp. 806–824, May 2007.

[14] R. Barra-Chicote et al., "Madrid-bsdm. http://lorien.die.upm.es/ barra/sintesis-albayzin08," Último acceso: septiembre de 2008.