DESCRIPCIÓN DEL SINTETIZADOR DE VOZ COTOVÍA PARA LA EVALUACIÓN ALBAYZIN TTS 2008

Eduardo R. Banga, Francisco Méndez, Francisco Campillo, Gonzalo Iglesias, Laura Docío

Grupo de Teoría de la Señal
Dpto. Teoría de la Señal y Comunicaciones
Universidad de Vigo – 36310 Vigo

RESUMEN

Este artículo describe el estado actual del sintetizador de voz basado en corpus Cotovía, desarrollado en la Universidad de Vigo con la colaboración del Centro Ramón Piñeiro para la Investigación en Humanidades. Cotovía es un sistema en el que se efectúa una búsqueda combinada tanto de las unidades acústicas y entonativas como de la estructura prosódica, con el objetivo de generar la voz sintética de mayor calidad posible a partir del corpus disponible.

1. INTRODUCCIÓN

Cotovía es un sistema de conversión texto–voz en gallego y castellano englobado dentro de las técnicas de concatenación. A diferencia de la mayoría de los sintetizadores de voz actuales, en los que se van generando las características fonéticas en una serie de etapas secuenciales, lo cual en cierta manera implica asumir independencia entre ellas, en Cotovía se aplica el concepto de la selección de unidades ([1]) tanto en la generación de la forma de onda como en el modelado entonativo, y se lleva un paso más allá escogiendo la mejor combinación de unidades acústicas y entonativas. De la misma forma, en la selección entonativa también se consideran diferentes estructuras entonativas, sacando así partido de la variabilidad de la voz, que permite que un mismo mensaje se pueda realizar de diferentes maneras sin afectar ni a la naturalidad ni a la inteligibilidad.

En este artículo se explican las características principales del sintetizador en el momento de presentarse a la evaluación Albayzin TTS 2008. En primer lugar, en la sección 2 se exponen los pasos que se siguieron para procesar la voz y poder generar a partir de ella la información necesaria para la síntesis. En La sección 3 se describen los principales módulos del sistema, desde la etapa lingüística hasta la generación de la forma de onda, incluyendo los diferentes modelos de estimación de la prosodia. Finalmente, en la sección 4 se presentan las conclusiones.

Este trabajo ha sido subvencionado por el Gobierno Español mediante el proyecto coordinado AVIVAVOZ (TEC2006-13694-C03)

2. GENERANDO LA VOZ

Para la evaluación de sistemas de conversión de voz Albayzin 2008 se ha puesto a disposición de los participantes el corpus upc_esma [2] como material de desarrollo. Este corpus consta de aproximadamente 1h 45 min. de voz (mono, frecuencia de muestreo 16 KHz, resolución de 16 bits por muestra), dividido en 3 subcorpus: frases fonéticamente equilibradas (506 ficheros, aproximadamente 30 minutos), párrafos fonéticamente equilibrados (208 ficheros, aproximadamente 30 minutos) y 45 minutos (62 ficheros) de párrafos literarios. Puesto que los niveles de grabación de cada subcorpus eran distintos, se ha hecho una normalización, fichero a fichero, al 70 % del valor máximo.

Para cada subcorpus se han proporcionado los ficheros de audio, la señal del laringógrafo, los ficheros de texto, la transcripción fonética (SAMPA) y una segmentación fonética revisada manualmente para una parte (todas las frases y 144 de los 208 párrafos fonéticamente equilibrados) y otra automática de la totalidad de los corpus.

Debido a discrepancias entre la transcripción y segmentación fonéticas proporcionados y el procesado lingüístico realizado por Cotovía, que es el que se utiliza para construir las voces de nuestro sistema, se ha decidido no utilizar directamente ninguna de las segmentaciones proporcionadas. En su lugar se ha adaptado de forma semi–automática a nuestro sistema la parte segmentada manualmente, realizan-

do una nueva segmentación automática del resto del material de desarrollo.

El proceso de segmentación automática [3] se ha realizado en dos etapas. En primer lugar, utilizando los ficheros segmentados de forma manual se ha entrenado un conjunto de HMMs continuos para cada una de las unidades fonéticas. Debido a la cantidad limitada de este conjunto de datos de entrenamiento se han considerado modelos de monofonemas independientes del contexto, con una topología de tres estados de izquierda–a–derecha y 4 gaussianas por estado. Con los modelos entrenados se ha realizado a continuación una segmentación automática de aquellos ficheros de los que no se dispone segmentación fonética manual. Dicha segmentación se ha realizado a través de un alineamiento forzado de Viterbi en el que se permite la posibilidad de insertar silencios/pausas opcionales entre palabras. El front–end utiliza como características 12 coeficientes melcepstrum, la log–energía, y sus correspondientes derivadas primeras y segundas.

Se han marcado de forma manual en todo el corpus de desarrollo las fronteras entonativas. El proceso automático de generación de voz para Cotovía ha requerido unas 7 horas de ejecución en un servidor Intel©Xeon^{TM}a 3.06 GHz con 2 GB de memoria RAM. Como herramientas externas se ha utilizado el programa Praat para calcular las marcas de pitch para la estimación de la frecuencia fundamental, el Festival para obtener la envolvente espectral (12 coeficientes MFCC) y el HTK para la segmentación automática.

3. DESCRIPCIÓN DEL SISTEMA

3.1. Módulo lingüístico

Como en cualquier aplicación de este estilo, el módulo lingüístico consta de una serie de fases en las que el texto de entrada se acaba transformando en una secuencia de unidades acústicas objetivo caracterizadas por un conjunto de factores que se emplean posteriormente en las etapas de modelado prosódico y generación de la forma de onda. En este caso, en Cotovía tiene especial relevancia la etapa de análisis morfosintáctico, ya no sólo por su importancia en la decisión del carácter tónico o átono de las palabras, sino por su influencia en la estimación de los contornos entonativos, tal y como se comenta en la sección 3.2.1. El analizador morfosintáctico empleado ([4]) consta de un conjunto reducido de reglas lingüísticas fiables, que eliminan para cada palabra aquellas categorías que no son posibles en función de su contexto, seguido de un analizador estadístico de ventana deslizante, en el que se decide la categoría más probable combinando un modelo contextual que considera la probabilidad de una secuencia de categorías, y otro modelo léxico, que considera la probabilidad de que una palabra tenga una categoría dada.

3.2. Estimación de la prosodia

Al igual que la mayoría de los sintetizadores de voz actuales, Cotovía incluye módulos de estimación de la energía, la duración, la entonación y de inserción de rupturas prosódicas. Sin embargo, en lugar de tratarse de una serie de módulos que se van ejecutando secuencialmente, en algunos de ellos se emplea la variabilidad de la prosodia para conseguir una mejor estimación conjunta. Así, por ejemplo, es el propio módulo entonativo el que se encarga de parte del problema de la inserción de rupturas prosódicas. A continuación se explica más detalladamente cada uno de los modelos.

3.2.1. Entonación

Cotovía emplea un modelo entonativo basado en corpus ([5]), con el grupo acentual (secuencia de palabras átonas que acaba en una palabra tónica), como unidad básica para la concatenación. Las principales características del modelo son:

- Cada grupo acentual se representa según su posición en el grupo fónico y entonativo, la posición en la frase, los tipos de frontera prosódica que lo rodean, el número de sílabas, la posición del acento, el tipo de oración (enunciativa, exclamativa, interrogativa e inacabada), la duración, la etiqueta morfosintáctica de la palabra tónica del grupo, el sintagma al que pertenece, y el sintagma que lo sigue.
- El coste de objetivo tiene en cuenta básicamente las desviaciones de las características antes mencionadas con respecto a los valores estimados. Lo más destacable es el tratamiento de la información gramatical ([6]), que se emplea tanto para penalizar la introducción o no de una ruptura entonativa entre dos grupos acentuales (ver sección 3.2.2), como para modelar el énfasis de los acentos.

- El coste de concatenación considera únicamente la continuidad de frecuencia fundamental y la continuidad de frontera prosódica (para evitar que se unan dos grupos que linden con diferentes fronteras en sus contextos originales).

3.2.2. Estructura prosódica

A diferencia de [5], donde se consideraban únicamente dos niveles de ruptura prosódica (pausa y no pausa), en la actualidad se consideran tres niveles de ruptura: pausa, no ruptura, y ruptura entonativa, definida ésta última como un límite de grupo entonativo que no coincide con pausa.

Las inserciones de pausas y de rupturas entonativas se tratan como dos problemas independientes. En primer lugar se decide la mejor posición para las pausas por medio de un árbol de clasificación, empleando factores como una ventana de cinco etiquetas morfosintácticas y la distancia en sílabas a las pausas circundantes. Posteriormente, en cada iteración del algoritmo de selección de unidades entonativas se consideran grupos acentuales candidato que pueden ir seguidos o no de una ruptura entonativa (en función del contexto del que fueron extraídos), tal y como se muestra en la Figura 1, donde los círculos sombreados representan grupos candidatos con una ruptura entonativa. De esta forma, modificando la función de coste de objetivo entonativo para considerar la inserción de rupturas entonativas (en este caso por medio de información sintáctica y morfosintáctica, tal y como se explica en [6]), el propio algoritmo de selección escoge la mejor combinación de grupos acentuales y estructura prosódica, produciendo una entonación sintética más variable y relacionada con el significado del mensaje que se desea transmitir. Como ejemplo de esta variabilidad, la Figura 2 muestra varios contornos posibles para la oración "El non sabía se saír ou quedar na casa" (*Él no sabía si salir o quedarse en casa*). Como se puede observar, el contorno 1 tiene una ruptura prosódica alrededor del instante $t \approx 0{,}8\ s$.

3.2.3. Duración

Por lo que respecta a la duración segmental, los fonemas se agrupan en diez clases (vocales abiertas, vocales medias, vocales cerradas, oclusivas sonoras, oclusivas sordas, fricativas sordas, laterales, nasales, vibrantes y silencio), y para cada una de ellas se calcula un modelo basado en regresión lineal multivariante, empleando como factores la identidad de los fonemas que lo rodean en una ventana de tamaño cinco, la posición en la palabra y en el grupo fónico, el tipo de oración y el carácter tónico o átono.

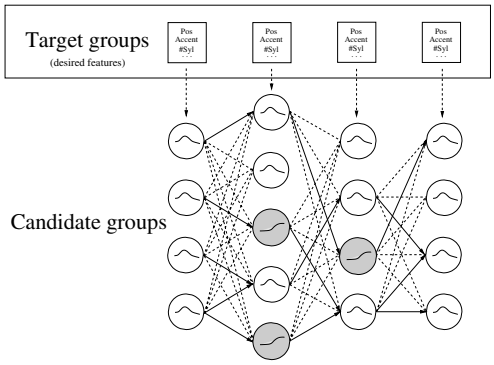

Figura 1. *Selección combinada de unidades entonativas y estructura prosódica ([6])*

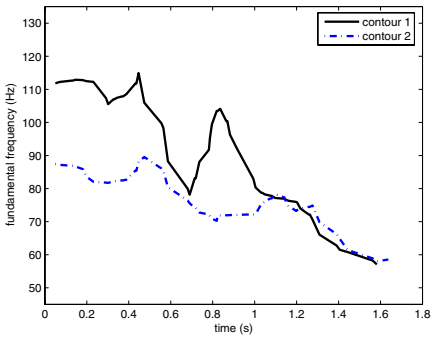

Figura 2. *Ejemplo de dos contornos con diferente estructura prosódica para una misma frase ([6])*

3.2.4. Energía

La energía es estimada por medio de un único modelo basado en regresión lineal multivariante, incluyendo como factores las clases del fonema y de los que lo rodean en una ventana de tamaño tres (según la misma clasificación del modelo de duración), la energía del fonema anterior, el número de sílabas desde el inicio y hasta el final del grupo fónico, y el carácter tónico o átono.

3.3. Selección de unidades acústicas

Tal y como sucedía con la entonación, en lo referente a las unidades acústicas Cotovía también esta basado en corpus, con el semifonema como unidad

básica para la concatenación. Además, dado que una misma frase se puede pronunciar con diferentes entonaciones sin afectar a su naturalidad, se repite la selección de unidades acústicas con cada uno de los N mejores contornos resultantes de la búsqueda entonativa ([5]), y se escoge la secuencia de semifonemas con mejor coste. Resumiendo, las principales características de la selección acústica son:

- Los semifonemas se parametrizan según su frecuencia fundamental, duración, energía, los fonemas que lo rodean, el carácter tónico, la posición en la palabra y en la frase, el tipo de frase a la que pertenece y los coeficientes cepstrales del semifonema y los que lo rodean.

- El coste de objetivo consta de dos partes ([7]). En primer lugar, la prosódica, donde se penalizan las desviaciones de la frecuencia fundamental, la duración y la energía con respecto a los valores predichos. Y en segundo lugar, la relacionada con la articulación del semifonema, donde se tienen en cuenta factores como los fonemas circundantes y la posición en la palabra y en la frase.

- El coste de concatenación considera la continuidad de frecuencia fundamental, energía y envolvente espectral.

3.4. Generación de la forma de onda

La señal sintética se genera mediante la concatenación de las formas de onda de las unidades acústicas escogidas. Cabe destacar que sólo se modifican prosódicamente aquellos semifonemas que se alejan de los valores estimados más de un umbral ($40\ ms$ para la duración y $5\ Hz$ para la frecuencia fundamental). Las unidades que no se tienen que modificar se copian directamente de la forma de onda original, recurriendo a las marcas de pitch únicamente en los puntos de concatenación.

4. CONCLUSIONES

En este artículo se ha descrito el estado actual del sintetizador de voz Cotovía, incluyendo tanto los pasos seguidos para la adición de una nueva voz, como el proceso que se sigue para la generación de la voz sintética. Durante la preparación de la voz para la evaluación quedó patente que pese a que la mayor parte del proceso es totalmente automático, es necesario desarrollar alguna herramienta que facilite el arduo proceso de revisión del etiquetado, sobre todo en lo referente a las fronteras prosódicas.

5. BIBLIOGRAFÍA

[1] A. Hunt y A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of ICASSP*, 1996, vol. 1, pp. 373–376.

[2] Antonio Bonafonte y Asuncion Moreno, "Documentation of the upc_esma spanish database," *TALP Research Center, Universitat Politecnica de Catalunya, Barcelona, Spain*, 2008.

[3] L. Docío y C. García, "Automatic segmentation of speech based on hidden markov models and acoustic features," in *Proceedings of 6th Internacional Conference on Spoken Language Processing*, 2000.

[4] F. Méndez, F. Campillo, E. R. Banga, y E. F. Rei, "Análisis morfológico estadístico en lengua gallega," *Procesamiento del lenguaje natural*, , no. 31, pp. 159–166, 2003.

[5] F. Campillo y E. R. Banga, "A method for combining intonation modelling and speech unit selection in corpus–based speech synthesis systems," *Speech Communication*, vol. 48, pp. 941–956, 2006.

[6] F. Campillo, J. van Santen, y E. R. Banga, "Combining phrasing and unit selection in intonation modelling," *IEE Electronic Letters*, vol. 44, no. 7, pp. 501–503, 2008.

[7] F. Campillo y E. R. Banga, "On the design of the cost functions for a unit selection speech synthesis," in *Proceedings of Eurospeech*, 2003, vol. 1, pp. 289–292.