Procesamiento del Lenguaje Natural, num. 29 (2002), pp.153-158 recibido 21-05-2002; aceptado 02-07-2002

Sistema de conversión texto-voz en lengua gallega basado en la selección combinada de unidades acústicas y prosódicas.

Eduardo Rodríguez Banga
Universidad de Vigo
ETSI Telecomunicación
erbanga@gts.tsc.uvigo.es

Francisco Campillo Díaz
Universidad de Vigo
ETSI Telecomunicación
campillo@gts.tsc.uvigo.es

Elisa Fernández Rei
Universidad de Santiago
Facultad de Filología
fgelisa@usc.es

Francisco Méndez Pazó
Universidad de Vigo
ETSI Telecomunicación
fmendez@gts.tsc.uvigo.es

Resumen: En esta comunicación se describe un sistema de conversión texto-voz en lengua gallega basado en las denominadas "técnicas de síntesis basadas en corpus". A diferencia de los tradicionales sintetizadores de voz por concatenación, que normalmente utilizan un conjunto de unidades de síntesis reducido, los sistemas de síntesis basados en corpus consideran múltiples realizaciones de cada unidad y, mediante técnicas de programación dinámica, seleccionan aquella secuencia de unidades que minimiza una función de coste. Por otro lado, tradicionalmente, la generación de la información prosódica se realiza en una etapa previa a la selección de unidades, lo que ocasiona que en muchas ocasiones sea necesario manipular en exceso las unidades seleccionadas con el fin de ajustarlas a la entonación, duración y energía deseadas. En este artículo también se propone la selección conjunta del contorno entonativo y de las unidades de síntesis, con objeto de minimizar la distorsión causada por las modificaciones prosódicas.

Palabras clave: síntesis de voz, conversión texto-voz, síntesis basada en corpus, entonación.

Abstract: In this contribution we describe a corpus-based text-to-speech system for Galician. While traditional concatenative speech-synthesis systems generally employ a quite reduced set of speech units, corpus-based synthesis systems consider many instances of every unit and, by means of dynamic programming techniques, select the sequence of units that minimizes a cost function. With reference to prosody, traditionally, the generation of the prosodic information is carried out in a previous stage to unit selection. This fact implies that, in many cases, the selected speech units must be manipulated in excess in order to fit the desired prosody. In this paper we also propose a method for combined selection of the intonation contour and the sequence of speech units in order to minimize the distortion due to prosodic modifications.

Keywords: speech synthesis, text-to-speech, corpus-based synthesis, intonation.

1 Introducción

En los últimos años los denominados sistemas de síntesis de voz basados en corpus han adquirido una gran relevancia debido a que son capaces de generar habla sintética con un alto grado de inteligibilidad y naturalidad. Este tipo de sistemas precisan de un amplio corpus de habla pregrabado a partir del que se seleccionan aquellas unidades que se consideran más adecuadas para la síntesis. Esta selección se realiza atendiendo a múltiples factores (identidad del alófono, contexto fonético, frecuencia fundamental y duración deseada, etc.) cuya influencia se refleja en una función de coste que, por supuesto, se trata de minimizar.

ISSN 1135-5948

Resulta evidente que los sistemas de síntesis basados en corpus necesitan una mayor capacidad de almacenamiento (memoria) que los sistemas basados en concatenación de un conjunto reducido de unidades. Mientras que el clásico sistema que considera una única realización de los distintos difonemas utiliza en torno a unas mil o dos mil unidades, los sistemas basados en corpus emplean la grabación completa (que suele ser de al menos 45 minutos o una hora). Por tanto, el número de realizaciones de cada unidad es claramente variable y realmente elevado en aquellas más frecuentes por lo que el algoritmo de selección de unidades será una de las etapas computacionalmente más costosas del sistema.

Hemos dejado hasta este punto la definición del tipo de unidad acústica (entendida como un segmento de voz pregrabado) utilizado en nuestro sistema de síntesis basado en corpus. Parecería lógico emplear unidades lo más largas posibles (en número de alófonos) con objeto de minimizar las posibles discontinuidades en las uniones entre unidades. No obstante, se ha comprobado la utilidad de unidades realmente pequeñas como los semifonemas siempre y cuando en la función de coste se prime la selección de semifonemas contiguos. En nuestro caso nos hemos decantado por esta última opción.

Una de las principales ventajas de los sistemas de síntesis basados en corpus reside en que al disponer de múltiples instancias de cada unidad se pueden considerar factores como su frecuencia fundamental, duración y energía en la etapa de selección. De esta forma, en gran parte de las ocasiones las modificaciones prosódicas que hay que realizar son pequeñas o, en ocasiones, incluso innecesarias. De esta forma se reduce la distorsión inherente a todo algoritmo de modificación prosódica.

En este trabajo también se plantea el utilizar técnicas similares a las empleadas durante la selección de las unidades acústicas para la obtención del contorno entonativo con el que se generará el habla sintética. En la primera versión de nuestro sistema de síntesis basado en corpus se seguía el método tradicional de generar el contorno entonativo en una etapa previa e independiente de la selección de unidades. El procedimiento seguido era la concatenación de patrones entonativos considerados representativos que eran asignados a los distintos grupos acentuales. Como consecuencia de este proceso, la entonación de la voz sintética resultaba bastante predecible y, por consiguiente, monótona, característica por otro lado compartida por gran parte de los sistemas de conversión texto-voz existentes.

En nuestra búsqueda de una mayor naturalidad en la voz sintética, a través de una mayor variabilidad en la evolución de la frecuencia fundamental, optamos por utilizar las técnicas de selección de unidades para la determinación del contorno entonativo. Decidimos considerar como unidad entonativa básica el contorno de frecuencia fundamental asignado a cada grupo acentual y, en lugar de decantarnos por un único contorno entonativo, proporcionar los N mejores candidatos (de acuerdo con una función de coste) para optar, finalmente, por aquel para el que se encuentre un conjunto de unidades acústicas (semifonemas) más apropiado. Se trata, por tanto, de generar una señal sintética lo más parecida posible a la voz del locutor original, al mismo tiempo que se logra una mayor variabilidad debido a que la secuencia de alófonos afecta a la selección del contorno entonativo.

En este artículo no describiremos todas las etapas que componen nuestro sistema de conversión texto-voz, etapas por otro lado ya clásicas (preprocesado del texto, silabación y acentuación, análisis morfosintáctico, trascripción fonética, etc.), sino que nos centraremos exclusivamente en los aspectos relativos a la síntesis basada en corpus. Por tanto, en primer lugar nos centraremos en la descripción del corpus grabado y en las técnicas de selección de las unidades acústicas para, posteriormente abordar el modelado prosódico y, muy especialmente, la generación del contorno entonativo.

2 Diseño del corpus

El corpus empleado para la selección de las unidades acústicas y de los patrones entonativos consta de unos 1300 enunciados, de los cuales unos 800 han sido diseñados de forma manual y 500 fueron extraídos automáticamente a partir de textos para considerar construcciones no

recogidas en el corpus manual, además de proporcionar algunas realizaciones más de construcciones ya consideradas.

El corpus manual recoge enunciados de siete modalidades oracionales: enunciativas, enumerativas, interrogativas (totales, parciales y alternativas), imperativas y exclamativas. Dentro de la modalidad enunciativa, se ha incluido un nutrido grupo de enunciados en los que se introdujo un elemento parentético (adverbial, modalizador, modificador explicativo, etc.) en el interior de la oración. Cada una de las oraciones está constituída por grupos acentuales, unidades prosódicas portadoras de un único acento que las caracteriza como grupo oxítono, paroxítono o proparoxítono. Los enunciados correspondientes a las diferentes modalidades combinan los diversos tipos de grupos acentuales y están constituídos por tres o más grupos acentuales, si bien también se incluyeron oraciones exclamativas e imperativas formadas por un único grupo acentual.

El corpus completo, así obtenido, fue grabado por un locutor profesional, escogido por tener una voz grave y agradable, siendo su duración real (excluyendo silencios) de poco más de una hora y estando constituido por unos 120.000 semifonemas.

3 Selección de unidades

3.1 Decisión del tipo de segmento acústico básico

Como ya se ha comentado anteriormente, en nuestro caso nos hemos decidido por la utilización de semifonemas, aunque considerando el contexto fonético izquierdo o derecho. Este tipo de unidad permite aprovechar al máximo la información del corpus grabado, y posibilita sustituciones sencillas en el caso poco probable de que no se encuentre la unidad con el contexto requerido. Aunque, inicialmente, este tipo de unidad parece tener el serio conveniente de que la concatenación no siempre ocurre en la zona estacionaria del alófono, este problema es enormemente paliado por la consideración de medidas de distancia espectral en la función de coste empleada en la selección de unidades, así como por el hecho de primar en la medida de lo posible el empleo de unidades consecutivas en el corpus. La calidad de la voz sintética obtenida con este procedimiento corrobora la idoneidad de los semifonemas como unidad de síntesis.

3.2 Organización de las unidades acústicas candidatas

La posibilidad de poder escoger entre un conjunto de candidatos semejantes para la síntesis de cada unidad acústica proporciona una flexibilidad no existente en los sistemas de concatenación no basados en corpus, pero tiene el inconveniente de aumentar de forma considerable la carga computacional del proceso.

En efecto, en la selección de unidades acústicas se combinan dos funciones de coste: función de coste objetivo y función de coste de concatenación. La primera mide la idoneidad de la unidad seleccionada atendiendo a factores como tonicidad, duración, frecuencia fundamental, etc., mientras que la segunda considera la distorsión resultante de concatenar las unidades seleccionadas. En principio, sería necesario calcular el valor de la función de coste objetivo para cada unidad candidata y, además, estimar el coste de concatenación para cada transición entre candidatos consecutivos. Empleando para la selección el conocido algoritmo de Viterbi, es relativamente sencillo comprobar que el número de funciones de coste que hay que evaluar es del orden de $\mathbf{N}^{*}n + (\mathbf{N} - 1)^{*}n^2$, siendo **N** el número de unidades de la frase y n el número medio de candidatos para cada unidad acústica. Así, por ejemplo, considerando una frase corta, de unas 60 unidades, y con un número medio de 200 unidades candidatas, sería preciso calcular unas 2.372.000 funciones de coste.

Queda patente, por tanto, la necesidad de organizar el corpus de unidades en una serie de grupos (**clusters**) en base a unas características que permitan reducir el número de unidades candidatas sin perder la flexibilidad antes mencionada. En nuestro caso, optamos por seguir unos criterios análogos a los descritos en (Febrer,2001) (Campbell y Black,1997) (Hunt y Black, 1996) , consistiendo en:

- Identidad del alófono.
- Contexto fonético: derecho o izquierdo.
- Tipo de proposición a la que pertenece el alófono: enunciativa, interrogativa, exclamativa o suspensiva.
- Posición en la frase: inicial (hasta la primera sílaba tónica, incluida), final (desde la última sílaba tónica, incluida) o media.
- Tonicidad silábica: tónica (si el alófono pertenece a una sílaba tónica) o átona.

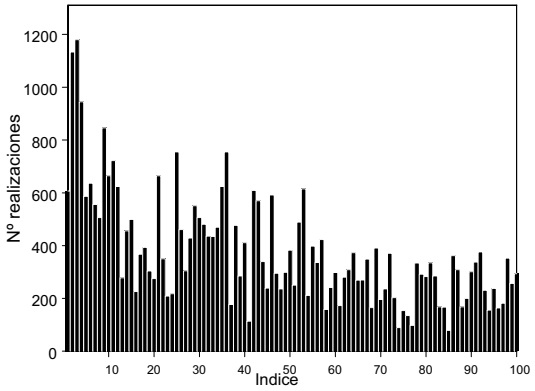

Figura 1. Número de realizaciones de los cien semifonemas más frecuentes del gallego (considerando el contexto fonético por la derecha).

A modo de ejemplo, el grupo {"#-a", izquierdo, enunciativa, inicial, átona} contiene todas las realizaciones del semifonema "a", con contexto por la izquierda "#" (silencio), pertenecientes a una sílaba átona situada al inicio de una frase enunciativa. De esta forma, el corpus de unidades queda dividido en una serie de grupos dentro de los cuales los segmentos acústicos se diferencian por sus características prosódicas (frecuencia fundamental, duración y energía). Para dar una idea de valores típicos, los grupos más frecuentes contienen varios centenares de realizaciones, mientras que los menos frecuentes disponen de unas pocas decenas. En la Figura 1 se representa el número de realizaciones de los cien semifonemas más frecuentes, teniendo en cuenta su contexto fonético derecho.

3.3 Búsqueda de la cadena óptima de unidades acústicas

En la Figura 2 se representa el procedimiento empleado para determinar la cadena de unidades acústicas más adecuada. Lo primero que llama la atención respecto a otros sistemas de conversión texto-voz es la consideración de varios contornos entonativos. La forma en que se generan estos contornos entonativos se tratará en la sección 4. La etapa de modelado prosódico también estima la duración de cada semifonema. El algoritmo de Viterbi se emplea para determinar la secuencia de unidades acústicas que minimiza la función de coste para cada uno de los posibles contornos. Este valor es posteriormente combinado (suma ponderada) con el de la función de coste empleada en la generación de los contornos entonativos. Cabe resaltar que como consecuencia de este proceso, no solamente se determina la secuencia de unidades acústicas, sino también el contorno entonativo.

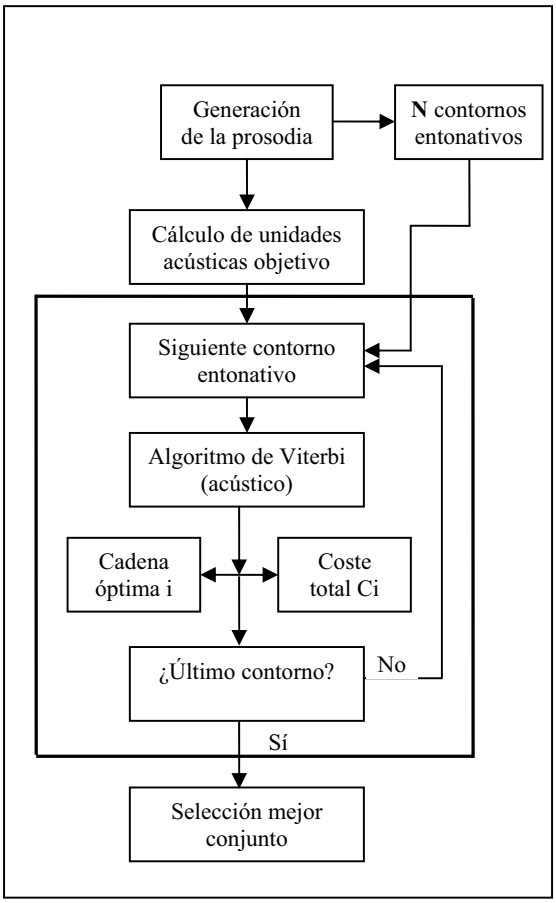

Figura 2. Selección de la secuencia de unidades.

3.4 Funciones de coste para la selección de la secuencia de unidades acústicas

El diseño de las funciones de coste es fundamental para el adecuado funcionamiento de un sistema de síntesis basado en corpus. Su importancia es, si cabe, todavía mayor cuando se utilizan unidades tan pequeñas como el semifonema ya que se deben seleccionar unidades de forma que no haya discontinuidades apreciables en los frecuentes puntos de unión. Una vez más, basándonos en trabajos de otros autores (Febrer,2001) (Campbell y Black, 1997), hemos decidido tener en cuenta los siguientes factores para la función de coste objetivo:

- Coste de palabra: se prima que los semifonemas pertenezcan a una misma palabra.
- Coste de trifonema: tiene en cuenta el contexto por el otro extremo del semifonema (que ya incluye el contexto izquierdo o derecho).
- Coste de duración: se busca que la duración del semifonema sea lo más próxima posible a la duración objetivo con objeto de evitar modificaciones excesivas al sintetizar.
- Coste de frecuencia fundamental: diferencia en la frecuencia fundamental al inicio y al final del semifonema respecto al valor determinado por el contorno entonativo. Una vez más se trata de evitar modificaciones prosódicas excesivas.
- Coste de tipo de proposición, posición del grupo acentual y posición de la sílaba tónica. Se emplean cuando no se encuentran unidades del tipo especificado y es necesario recurrir a unidades consideradas "similares".

Durante el cálculo de la función de coste objetivo se marcan aquellas unidades cuya duración y frecuencia fundamental se alejan de los valores deseados una distancia mayor que unos determinados umbrales (por ejemplo, 10 ms para la duración y 5 Hz para la frecuencia fundamental), para modificarlas posteriormente prosódicamente en caso de que fuesen escogidas. En caso contrario, no se realiza modificación prosódica.

En la función de coste de concatenación se potencia la continuidad espectral de los semifonemas, mediante la utilización de una medida de distancia cepstral, y la continuidad de energía. Cabe destacar que esta última consideración es especialmente importante en sistemas que, como el nuestro, no modelan explícitamente el contorno de energía.

4 Modelado prosódico

Como ya comentamos con anterioridad, en la etapa de modelado prosódico se realiza la estimación de la duración de los semifonemas y la generación de los N contornos entonativos candidatos. La predicción de las duraciones segmentales se realiza mediante las clásicas técnicas de regresión múltiple, considerándose factores como la identidad del alófono, contexto fonético, número de sílabas desde el principio del grupo fónico y hasta el final, etc. No obstante en este apartado nos centraremos en la generación de los contornos entonativos por ser la aportación más novedosa.

En nuestro sistema el patrón entonativo asociado al grupo acentual (entendido como una secuencia de palabras átonas que finaliza en una palabra tónica) se considera la unidad entonativa básica. Una vez definida la unidad básica, la generación de los contornos entonativos es un problema formalmente equivalente a la selección de la secuencia de unidades acústicas (Malfrère, Dutoit y Mertens, 1998), con la salvedad de que es necesario el diseño de unas funciones de coste (objetivo y de concatenación) adecuadas para esta tarea. En primer lugar, se organizan los grupos acentuales en conjuntos en función del tipo de proposición (con la misma clasificación que las unidades acústicas), la posición en la frase (inicial, media o final), y la palabra que lleva el acento (aguda, llana o esdrújula). Posteriormente, se escoge el más próximo dentro del conjunto al que pertenezca el grupo que se desea sintetizar. Los factores que actualmente consideramos para la función de coste objetivo son los siguientes:

- Número de sílabas del grupo acentual.
- Duración temporal del grupo acentual.
- Posición porcentual del grupo entonativo en la frase.
- Posición porcentual del grupo acentual en el grupo entonativo.

En cuanto a la función de coste de concatenación, consideramos:

- Continuidad de frecuencia fundamental en las uniones entre las unidades entonativas.
- Evolución temporal de la frecuencia fundamental media de las unidades entonativas teniendo en cuenta el tipo de proposición (enunciativa, interrogativa...).
- Diferencia entre los valores máximos de frecuencia fundamental de unidades entonativas consecutivas.

Estos dos últimos factores de la función de coste tratan de modelar, aunque de forma muy flexible, la declinación. Por supuesto, a aquellos patrones entonativos asociados a grupos acentuales consecutivos en el corpus grabado se les asigna un coste de concatenación nulo.

Las unidades entonativas empleadas durante el proceso de selección han sido también extraídas a partir del corpus grabado. Los contornos de frecuencia fundamental de las frases del corpus fueron estimados, interpolados en los segmentos sordos y suavizados utilizando el programa Praat (Boersma y Weenink, 2001). En la Figura 3 se presenta un histograma que muestra la cobertura del corpus en términos del número de grupos acentuales por grupo fónico. Evidentemente, son más frecuentes los grupos fónicos con pocos grupos acentuales, aunque también hay una cobertura razonable de grupos fónicos largos.

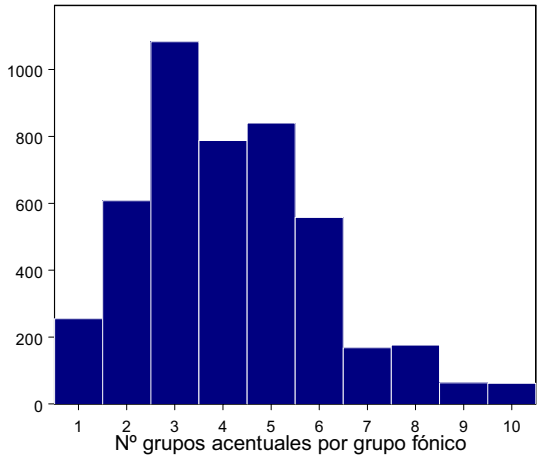

Figura 3. Cobertura del corpus en términos del número de grupos acentuales por grupo fónico

5 Conclusiones

En este artículo hemos mostrado como en un sistema de síntesis de voz basado en corpus se pueden combinar la etapa de selección de las unidades acústicas y la etapa de generación del contorno entonativo. El fin perseguido es conseguir que la voz sintética se parezca lo más posible a la voz del locutor original. Este fin se consigue considerando inicialmente varios contornos entonativos, para después decantarnos por aquél para el que existe una secuencia de unidades más apropiada (aunque teniendo también en cuenta el coste de cada contorno entonativo candiadato). De esta forma se reduce la distorsión debida a manipulaciones prosódicas excesivas, al mismo tiempo que se obtiene una mayor riqueza en los contornos entonativos.

Bibliografía

Campbell N. & Black A.W., "Prosody and the Selection of Source Units for Concatenative Synthesis" chapter in "Progress in Speech Synthesis". Eds J. van Santen, R Sproat, J Olive and J. Hirschberg, pp 279-282, Springer Verlag.1997

Hunt A.J. & Black A.W., "Unit Selection in a Concatenative Speech Synthesis using Large Speech Database", Proceedings of ICASSP96, pp. 373-376, 1996.

Febrer, A., "Síntesi de la parla per concatenació basada en la selecció". Tesi doctoral. Departament de Teoria del Senyal i Comunicacions. Universitat Politécnica de Catalunya. 2001

Malfrère, F.; Dutoit, T. & Mertens, P. "Automatic prosody generation using suprasegmental unit selection".Proc. 3rd ESCA/COCSADA Workshop on Speech Synthesis, Jenolan Caves, Australia (December 1998), pp. 323-328

Boersma P. & Weenink D. "Praat: doing phonetics by computer". http://www.praat.org. 2001