NUEVO MÓDULO DE ANÁLISIS PROSÓDICO DEL CONVERSOR TEXTO-VOZ MULTILINGÜE DE TELEFÓNICA I+D

M. Á. Rodríguez[1], J. G. Escalada[1], A. Armenta[1] y J. M. Garrido[2]

[1]División de Tecnología del Habla
Telefónica Investigación y Desarrollo
Emilio Vargas, 6, 28043 Madrid
Via Augusta, 177, 08021 Barcelona

[2]Departament de Traducció i Filologia, UPF/
Barcelona Media Centre d'Innovació
Ocata, 1, 08003 Barcelona

RESUMEN

Este artículo presenta un nuevo módulo de análisis prosódico para un conversor texto-voz (CTV) que se ocupa de predecir y caracterizar los límites prosódicos en la lectura de un texto. Los límites tratados son tanto pausas como frases entonativas, y se emplean para mejorar la generación de otros parámetros prosódicos (duración de los sonidos y contorno de F0) y aumentar la naturalidad de la voz sintética. El funcionamiento del módulo de análisis prosódico no sólo tiene en cuenta características lingüísticas generales propias de un idioma determinado, sino que también se adapta al modo particular de hablar de un locutor humano de referencia.

1. INTRODUCCIÓN

En este artículo se presenta una evolución del módulo de inserción de pausas (módulo estructurador-pausador) del CTV multilingüe de Telefónica I+D [1], orientada a mejorar tanto la naturalidad y corrección de las pausas generadas, como la generación del resto de parámetros prosódicos. El antiguo módulo pausador sólo consideraba límites de grupo fónico (pausas) que se decidían por regla sobre el resultado de una agrupación en sintagmas. Las mejoras se han centrado fundamentalmente en dos aspectos: por un lado, la detección de pausas se ha enriquecido con la inserción de otro tipo de límites prosódicos, con lo que ya no cabe hablar tanto de un módulo pausador sino de un módulo de análisis prosódico, encargado de identificar en los textos de entrada unidades prosódicas de distinto ámbito; y por otro, se ha intentado modelar la variación estilística interlocutor que se da en la segmentación prosódica de los enunciados.

Empieza a ser un hecho generalmente aceptado que los enunciados de las lenguas presentan una estructura prosódica más compleja que la simple segmentación en grupos fónicos o entonativos. Desde hace tiempo se han propuesto otras unidades prosódicas de nivel inferior, como el grupo acentual [2, 3], el grupo tónico [4] o la frase intermedia [5, 6]. Estas unidades formarían toda una estructura jerárquica que determinaría, entre otros aspectos, la asignación de las pausas y de los contornos entonativos [7, 8, 9]. Este tipo de aproximación jerárquica está presente ya en algunas implementaciones de analizadores prosódicos para conversión texto-voz [10, 11].

De acuerdo con esta idea, en el módulo de análisis prosódico del CTV de Telefónica se han distinguido dos tipos distintos de límites prosódicos: límites de nivel 1, que se corresponderían con los límites de grupo fónico o entonativo en la terminología lingüística tradicional, y que se realizan en el nivel fonético con la inserción de una pausa de una duración específica y un movimiento de F0 determinado; y límites de nivel 2, de ámbito inferior al anterior, más o menos equivalentes a las frases intermedias del modelo autosegmental (aquí proponemos la denominación frase entonativa), y que se manifiestan fonéticamente con un movimiento específico de F0, pero sin inserción de pausa.

La asignación de límites prosódicos de nivel 1 y 2 se realiza teniendo en cuenta tres factores: organización sintáctica de los enunciados (*chunks*); información sobre el número de sílabas desde el último límite; e información sobre la velocidad de elocución. La organización sintáctica y el número de sílabas ya se empleaban en el antiguo módulo pausador (si bien de una manera mucho menos elaborada que en la versión actual); la velocidad de elocución se ha incorporado en esta nueva versión.

Por otro lado, el antiguo módulo pausador no admitía variación interlocutor en el pausado. Todos los locutores sintéticos de un mismo idioma pausaban exactamente igual el mismo texto. En el nuevo módulo de análisis prosódico se ha intentado recoger la variación interlocutor, mediante la creación de modelos específicos para cada locutor sintético. Los datos sobre la variación estilística propia del locutor para la construcción de sus modelos se extraen del análisis prosódico del corpus grabado por cada locutor de referencia. La creación de modelos para la segmentación prosódica a partir del análisis automático

de datos reales no es nueva en conversión texto-voz ([12], por ejemplo), aunque en este caso se ha aplicado a la creación de modelos específicos para cada locutor.

Este nuevo procedimiento de predicción de límites prosódicos se comenzó a aplicar a los locutores desarrollados para español castellano, pero luego se ha aplicado también a los locutores del resto de los idiomas incorporados en el CTV multilingüe de Telefónica.

2. ESTRUCTURA Y FUNCIONAMIENTO DEL MÓDULO DE ANÁLISIS PROSÓDICO

En la Figura 1 se incluye un diagrama que ilustra el proceso completo que sigue el módulo de análisis prosódico.

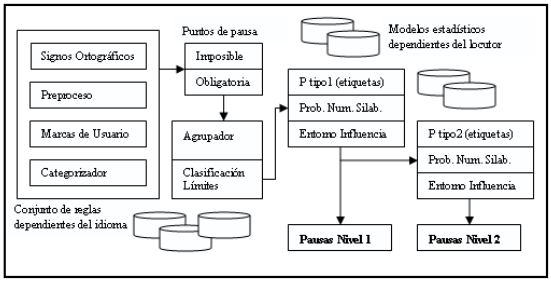

Figura 1. *Diagrama del módulo de análisis prosódico*

De forma simplificada, el proceso de segmentación del texto de entrada en unidades prosódicas implica dos fases:

- La identificación y etiquetado de los puntos en el texto que son candidatos a recibir una marca de límite (límites potenciales). Entre ellos están los signos ortográficos (se considera que no siempre implican pausa).
- La selección de los límites definitivos entre los límites potenciales.

Para la selección definitiva de los límites se utiliza, además del conjunto de límites potenciales ya etiquetados y caracterizados, una serie de datos sobre el comportamiento prosódico de cada locutor de referencia, extraídos del análisis automático del corpus grabado. Estos datos son los que permiten obtener la variación interlocutor en el proceso de pausado automático.

En los siguientes apartados se describen con más detalles estos dos procesos.

2.1. Detección y clasificación de los límites potenciales

Para la localización de los límites potenciales se lleva a cabo en el CTV un análisis morfológico del texto, como paso previo a la agrupación de palabras en sintagmas *(chunking)*. Estas tareas se realizan en los módulos categorizador y agrupador. El primero es el encargado de asignar una categoría a las palabras del texto, y de determinar sus propiedades morfológicas (género, número, etc.), y el segundo se ocupa de agrupar las palabras categorizadas en algo parecido a sintagmas (pseudosintagmas). Este análisis se lleva a cabo mediante un sistema de reglas y de diccionarios específicos para cada idioma.

El siguiente paso es caracterizar los límites potenciales con una etiqueta que se empleará para el proceso de selección de los límites definitivos. Al igual que en el caso del análisis morfosintáctico, esta tarea la realiza un conjunto de reglas y diccionarios específicos de cada idioma, que en función del contexto sintáctico asignan una etiqueta a cada límite potencial identificado. Se han definido más de 170 etiquetas diferentes para el etiquetado de los límites, que intentan reflejar, fundamentalmente, el contexto sintáctico del límite potencial.

2.2. Selección de los límites definitivos

El proceso de selección de límites se ha organizado en dos fases secuenciales. En la primera fase se deciden los límites definitivos de nivel 1 entre todos los límites potenciales. En la segunda fase, se seleccionan los límites definitivos de nivel 2 entre los límites potenciales restantes.

Ambas fases funcionan de una forma semejante, si bien emplean datos diferentes durante su proceso.

2.2.1. Selección de límites definitivos de nivel 1

Los datos del locutor humano de referencia que se emplean en esta fase son:

• Probabilidades de realizar límite de nivel 1 para cada tipo (etiqueta) de límite. Reflejan el hecho de que las distintas etiquetas de los límites potenciales no tienen la misma probabilidad de inducir un límite prosódico. Estas probabilidades se obtienen encontrando el número de veces que cada etiqueta ha sido realizada por el locutor de referencia como límite de nivel 1, y dividiendo por el número de ocurrencias de esa etiqueta en el corpus de grabación.

• Probabilidad acumulada por número de sílabas transcurridas. Esta tabla de valores refleja el hecho de que a medida que aumenta el número de sílabas desde un límite de nivel 1, aumenta la probabilidad de introducir un nuevo límite de nivel 1. Para su cálculo, primero se obtiene la probabilidad de encontrar en el corpus grabado grupos fónicos (delimitados por límites de nivel 1) de determinado número de sílabas (una, dos, tres...). Una vez asociadas las probabilidades a cada número de sílabas, se acumulan para obtener una especie de función de densidad de probabilidad (probabilidad de que un grupo fónico tenga x sílabas o menos). Al hacer este cálculo no se tienen en cuenta los límites asociados a signos ortográficos, pues se considera que su realización como límites definitivos no está tan condicionada por la longitud de los grupos fónicos.

Respecto a las probabilidades asociadas a cada etiqueta de límite, es bien conocido que la bondad de determinada etiqueta para ser escogida como límite definitivo no depende únicamente de su identidad sino también del contexto en el que se encuentra. El volumen de la combinatoria entre distintos códigos de pausa hace inviable obtener valores de probabilidad para todos los casos de combinaciones. Por esta razón se ha buscado una aproximación alternativa que permita tener en cuenta la influencia del contexto en la realización de un límite como definitivo. Esta alternativa consiste en modificar los valores de probabilidad inicial asociados a cada etiqueta, resultando unas probabilidades incentivadas (o penalizadas, si el incentivo resulta negativo). Para cada límite potencial se localizan los límites anterior y siguiente cuyas probabilidades iniciales sean mayores que las del límite dado. Esto define un entorno de influencia del código de pausa. Cuanto más extenso sea el entorno de influencia, más preferible será la pausa. El incentivo (o penalización) de probabilidad se calcula a partir de los datos de probabilidades acumuladas por número de sílabas transcurridas.

Una vez obtenidas las probabilidades incentivadas correspondientes a cada límite, se pasa a la etapa final del procedimiento de selección, que consiste en un algoritmo de programación dinámica tipo Viterbi que obtiene la secuencia óptima de valores seleccionado / no-seleccionado correspondiente a la secuencia de límites potenciales de una frase de texto. En cada punto del camino (cada límite) se obtienen dos valores de suma de probabilidad: uno correspondiente al caso de seleccionar ese límite, y otro correspondiente al caso de no seleccionarlo. En el punto final del camino se escoge la alternativa de mayor suma de probabilidades acumulada, se va siguiendo de qué alternativa del límite anterior procede (seleccionado / no-seleccionado), y se reconstruye hacia atrás el camino óptimo.

2.2.2. Selección de límites definitivos de nivel 2

La selección de límites de nivel 2 sigue el mismo procedimiento descrito anteriormente, teniendo en cuenta que se parte de los límites de nivel 1 ya seleccionados. En este caso, los datos que se emplean procedentes del análisis del corpus de grabaciones son los siguientes:

• Probabilidades de realizar límite de nivel 2. Se calculan encontrando el número de veces que cada etiqueta ha sido realizada por el locutor de referencia como límite de nivel 2, y dividiendo por el número de ocurrencias de esa etiqueta en el corpus de grabación, sin tener en cuenta los casos en que esa etiqueta se realizó como límite de nivel 1.

• Probabilidad acumulada por número de sílabas transcurridas. Para su cálculo, primero se obtiene la probabilidad de encontrar en el corpus grabado frases entonativas (limitadas por la izquierda por un límite de nivel 1 o de nivel 2, y limitadas por la derecha por un límite de nivel 2) de determinado número de sílabas (una, dos, tres,...). Como en el caso de los límites de nivel 1, una vez asociadas las probabilidades a cada número de sílabas, se acumulan para obtener una especie de función de densidad de probabilidad.

3. EVALUACIÓN

El verdadero valor de la predicción y caracterización de límites prosódicos está en la mejora que aporta en la corrección y naturalidad en la lectura de textos por parte del CTV. Desde este punto de vista, la mejor forma de evaluación sería mediante pruebas subjetivas con ejemplos de voz sintetizada, que tendrían que ser escuchadas y evaluadas por un número suficiente de personas para expresar y cuantificar sus preferencias. Estas pruebas son complejas y costosas. Además, los cambios introducidos en nuestro sistema han llevado aparejados otros cambios que afectan a los módulos generadores de otros parámetros prosódicos (duración y F0), por lo que la evaluación subjetiva no permitiría discernir el efecto individual del nuevo módulo de análisis prosódico. Por último, cuando se habla de estructura prosódica, el concepto de corrección es algo que, en gran medida, está por definir, y seguramente también sometido a variación estilística.

Por ello, hemos optado por realizar una evaluación de tipo objetivo. Asumiendo como referencia de corrección la segmentación prosódica realizada por los locutores de referencia en la lectura del corpus. Esta evaluación tiene sus limitaciones, pero nos permite comprobar si el módulo de análisis prosódico consigue sus objetivos: realizar una correcta predicción de límites prosódicos, y reflejar las peculiaridades de un hablante determinado en esta tarea. Hemos tomado como referencia las grabaciones realizadas por dos locutores masculinos en español castellano (identificados como JOSÉ y NACHO), que presentan una forma bastante diferente de realizar límites prosódicos al leer textos.

Se han escogido 10 oraciones con suficiente longitud como para inducir un buen número de límites internos de nivel 1 (pausas) y de nivel 2 (frases entonativas), y con una estructura compleja y variada. El conjunto de esas frases contiene un total de 292 lugares en los que se podría introducir un límite (espacios entre palabras), descontando las pausas finales correspondientes a cada oración.

Para comparar el funcionamiento del módulo de análisis prosódico ajustado a cada uno de los dos locutores de referencia, presentamos las tablas 1 y 2, que recogen los siguientes datos: lugares en los que tanto el locutor de referencia como su correspondiente locutor sintético coinciden en no hacer ningún límite (CNLI); coincidencia total (lugar y nivel) en hacer un límite (CTOT); coincidencia parcial (coincide el lugar pero no el nivel) en hacer un límite (CPAR); límites realizados por el locutor pero no predichos por el sistema (NPRE); y límites predichos por el sistema pero no realizados por el locutor (falsas predicciones, FPRE).

CNLI	CTOT	CPAR	NPRE	FPRE	Total
197 casos	37 casos	16 casos	33 casos	9 casos	292 casos
67,47%	12,67%	5,48%	11,3%	3,08%	100%

Tabla 1. *Comparación de la asignación de límites del locutor sintético JOSÉ con su correspondiente locutor humano de referencia teniendo en cuenta todas las posibilidades de hacer límite en los textos*

CNLI	CTOT	CPAR	NPRE	FPRE	Total
219 casos	39 casos	7 casos	16 casos	11 casos	292 casos
75%	13,36%	2,4%	5,48%	3,77%	100%

Tabla 2. *Comparación de la asignación de límites del locutor sintético NACHO con su correspondiente locutor humano de referencia teniendo en cuenta todas las posibilidades de hacer límite en los textos*

CNLI y CTOT recogen el funcionamiento estrictamente correcto (JOSÉ: 80,14%, NACHO: 88.36%). CPAR recoge un margen añadido de funcionamiento aceptable (JOSÉ: 5,48%, NACHO: 2,4%). NPRE y FPRE se pueden considerar errores de funcionamiento, aunque no se refieren a límites objetivamente mal ignorados o mal predichos por el sistema. La lectura resultante puede ser correcta o al menos aceptable para el texto de entrada, pero los hemos considerado errores de funcionamiento en tanto en que se apartan de lo que hizo el locutor de referencia

La diferencia de comportamiento entre los dos locutores de referencia frente a los mismos textos aparece en la tabla 3, donde se recogen los límites que realizaron: de nivel 1, distinguiendo los relacionados con signos ortográficos (N1OR) y los no relacionados (N1NO), y los de nivel 2 (NIV2). También se muestran las coincidencias totales (lugar y nivel) entre los dos locutores (CTOT). Se puede ver que el número total de límites realizados por uno y otro cambia bastante, y que el grado de coincidencia es relativamente bajo, excepto en el caso de las pausas relacionadas con signos ortográficos.

	N1OR	N1NO	NIV2	Total
JOSÉ	18	15	53	86
NACHO	20	13	29	62
CTOT	18	8	17	43

Tabla 3. *Diferencias de comportamiento entre los locutores de referencia JOSÉ y NACHO*

4. CONCLUSIONES

En este trabajo se ha presentado un nuevo módulo de análisis prosódico para el CTV de Telefónica. Además de mejorar la corrección del pausado automático con respecto a la versión anterior, el principal objetivo ha sido hacer la tarea de segmentación prosódica automática dependiente del locutor, con la idea de obtener resultados distintos para un mismo texto en función del locutor sintético que se utilice para leerlo.

Los resultados de la evaluación llevada a cabo muestran un grado aceptable de funcionamiento del sistema, adaptado a distintos locutores. Además, en una evaluación subjetiva informal de la voz generada por el CTV, la voz sintética resulta más natural y, en particular, menos monótona y predecible en su entonación. También se ha podido apreciar que la inserción de límites con distintos locutores sobre el mismo texto no coincide en muchos casos.

10. BIBLIOGRAFÍA

[1] M. Á Rodríguez, J. G. Escalada y D. Torre, "Conversor texto.voz multilingüe para español, catalán, gallego y euskera", Procesamiento del Lenguaje Natural, Revista nº 23 SEPLN, pp. 16-23, 1998.

[2] N. Thorsen, "Interpreting Raw Fundamental-Frequency Tracings of Danish", Phonetica, 36, pp. 57-78, 1979.

[3] N. Thorsen, "Intonation contours and stress group patterns in declarative sentences of varying length in ASC Danish", ARIPUC 14, pp. 1-29, 1980.

[4] T, Navarro, *Manual de entonación española*, Guadarrama, Madrid. 1944.

[5] M. Beckman y J.B. Pierrehuumbert, "Intonational structure in English and Japanese", Phonology Yearbook, 3, pp. 255-310, 1986.

[6] P. Prieto, "Phonological phrasing in Spanish" en: Optimality-Theoretic Advances in Spanish Phonology, ed. by Sonia Colina and Fernando Martínez-Gil, pp. 39-60. John Benjamins, Amsterdam/Philadelphia, 2006.

[7] E. O Selkirk*, Phonology and Syntax*: *The relation between Sound and Structure*, Cambridge, MA, The MIT Press, 1984.

[8] M. Nespor y I. Vogel,. *Prosodic Phonology*, Dordrecht, Foris, Studies in Generative Grammar, 28, 1986.

[9] D.R. Ladd, "Intonational phrasing: the case for recursive prosodic structure", Phonology Yearbook, 3, 311-340, 1986.

[10] B. Gili Fivela y S. Quazza, "A Prosodic Parser for an Italian Text-to-Speech System", Actas del XII Congreso de la SEPLN, Sevilla, septiembre de 1996. Procesamiento del Lenguaje Natural, Revista 19: pp. 189-200, 1996.

[11] B. Gili Fivela y S. Quazza, "Text-to-prosody parsing in an Italian synthesizer", in Proceedings of the 5th European Conference On Speech Communication and Technology (EuroSpeech), pp. 987-990, 1997.

[12] J. Hirschberg. y P. Prieto, "Training intonational phrasing rules automatically for English and Spanish text-to-speech", Speech Communication 18,3, pp. 283-292, 1996.