AMIGO: UN CONVERSOR TEXTO-VOZ PARA ESPAÑOL

Miguel _ngel Rodríguez Crespo
José Gregorio Escalada Sardina
Alejandro Macarrón Larumbe
Luis Monzón Serrano

Telefónica, I+D
Emilio Vargas nº6
28043 MADRID
e-mail: miguel@craso.tid.es

1 INTRODUCCIÓN

Las actividades de la división de Tecnología del Habla de Telefónica I+D se orientan a la introducción de nuevos servicios en la red telefónica y a hacer más fáciles de usar servicios ya existentes, añadiéndoles nuevas prestaciones.

El lenguaje hablado es la forma de comunicación más habitual entre los seres humanos, pero hoy día es cada vez más frecuente que nos *comuniquemos* con ordenadores, y la comunicación con ellos suele hacerse mediante equipos (terminales) que no son cómodos de usar por todo el mundo y a los que no se puede acceder desde cualquier sitio.

Por otra parte, el teléfono es un medio de comunicación que es fácil de usar y que se encuentra ampliamente difundido.

Por tanto, sería muy útil poder comunicarnos con un ordenador usando un teléfono convencional y empleando el lenguaje hablado, pudiendo así acceder a diversas informaciones almacenadas en el ordenador de manera sencilla y cómoda. Pero para ello sería necesario que los ordenadores *oyeran* y *hablaran*.

Con estos propósitos, en la división de Tecnología del Habla de Telefónica I+D se trabaja principalmente en dos campos: reconocimiento de voz y conversión texto-voz.

Es evidente que las aplicaciones de las tecnologías del habla no se limitan a lo dicho anteriormente. Piénsese, por ejemplo, lo importante que sería para personas con algún tipo de incapacidad física (ciegos, sordos, mudos, paralíticos, ...) disponer de máquinas que obedecieran a su voz y/o que respondieran hablando.

En este artículo vamos a presentar el conversor texto-voz AMIGO, que ha sido desarrollado dentro de la división de Tecnología del Habla de Telefónica I+D.

2 DIVISIÓN EN BLOQUES

La conversión automática de texto a voz supone una serie de procesos de tipo muy diverso.

Por un lado, es necesario realizar una serie de tareas de tipo lingüístico que, partiendo del análisis del texto de entrada, puedan proporcionar datos útiles para la correcta y natural *lectura* del texto.

Por otro lado, es necesario generar una señal de voz mediante métodos electrónicos. Esto supone el empleo de técnicas de proceso digital de voz.

En el conversor texto-voz AMIGO, estas diferentes tareas han sido separadas y se corresponden con dos grandes bloques, de los que hablaremos en los dos siguientes apartados.

3 BLOQUE DE PROCESO LINGÜÍSTICO

Para leer un texto es preciso saber qué sonidos hay que producir, y cómo producirlos. Estos son los objetivos de este módulo, que se traducen en obtener la cadena de alófonos correspondiente al texto de entrada e información de prosodia referente a la producción de los mismos: duración de los alófonos y evolución temporal de la frecuencia fundamental (F0) a lo largo del discurso.

Esta información no se generará de una vez para todo el fichero de texto completo, pues el conversor debe ir generando voz a medida que procesa el texto. Si hubiera que esperar hasta procesar todo el texto, se tardaría bastante tiempo (dependiente del tamaño del fichero) en empezar a producir la voz sintética.

Por tanto, se ha definido una unidad de trabajo para el proceso lingüístico, a la que hemos llamado *frase* (que no siempre coincidirá realmente con lo que se entiende por frase).

Sobre una frase se efectúan distintas tareas individuales que van generando información intermedia. El conjunto de estas tareas han sido agrupadas en módulos, de los que se hablará más adelante. Después de procesar una frase, se pasará a procesar la siguiente. Se supone que las frases son unidades independientes, por lo cual los resultados de procesar una frase no deben afectar al proceso de otras frases.

Las informaciones que se generan durante el proceso lingüístico se recogen en diversas estructuras de datos, que son compartidas por todos los módulos. Las más importantes son las siguientes:

* **Vector frase:** recoge la secuencia de caracteres del texto de entrada que se ha decidido que es una frase.

* **Vector de palabras:** cada uno de los elementos del vector es una estructura en la que se recogen distintos datos concernientes a cada una de las *palabras* de la frase. Hacemos notar que una *palabra* no siempre coincidirá con lo que se entiende por palabra. Los datos a los que nos hemos referido anteriormente son:

  - Secuencia de caracteres que componen la palabra.

  - Código que identifica la naturaleza de la palabra (por ejemplo, si es un signo ortográfico, una palabra sólo compuesta de letras, ...).

  - División en sílabas de la palabra, y posición del acento fonético (si lo hay).

  - Código que indica el modo de escritura que originalmente tenía la palabra (por ejemplo, todas las letras en mayúsculas, inicial en mayúsculas y resto en minúsculas, ...).

  - Código que indica la categoría gramatical asignada a la palabra, y alguna información adicional (por ejemplo, si forma parte de una perífrasis verbal).

  - Código con información acerca del género y del número de la palabra. Esta información sólo se generará en ciertos casos.

  - Código con información acerca del tipo de acentuación fonética que tiene la palabra (definitivamente acentuada, definitivamente inacentuada, acentuación no decidida todavía).

- Número de acentos que tiene la palabra (dos, uno o ninguno).

- Número de sílabas que tiene la palabra.

* **Estructura de nodos del análisis sintáctico:** recoge información acerca de la estructura sintáctica de la frase. Cada uno de los nodos de la estructura se corresponde con un sintagma de la frase y contiene los siguientes campos de información:

  - Número de sílabas del sintagma.

  - Número de sílabas tónicas del sintagma.

  - Número de sintagmas que se encuentran inmediatamente por debajo del sintagma considerado (*descendientes directos*).

  - Código que indica el tipo de sintagma (sujeto, predicado, ...).

  - Índice de la palabra en la cual comienza el sintagma.

  - Índice de la palabra en la cual finaliza el sintagma.

  - Una serie de punteros que señalan a cada uno de los nodos que corresponden a cada uno de los descendientes directos.

  - Un puntero que señala al sintagma que se encuentra inmediatamente por encima del considerado (*ascendiente directo*).

* **Vector de códigos de pausa:** indica para cada una de las palabras de la frase si hay que realizar una pausa antes de ella o no. En caso de que haya que realizar una pausa, el código indica el tipo de pausa que hay que realizar.

* **Vector de alófonos:** cada una de sus posiciones guarda distintos datos relativos a cada uno de los alófonos de la frase. Estos datos son:

  - Representación alfabética del alófono, según un conjunto de alófonos amplio.

  - Representación alfabética del alófono, según un conjunto de alófonos más reducido que el anterior. Más adelante se explicarán las razones de esta doble representación.

  - Código numérico equivalente del alófono.

  - Indicación acerca de si el alófono está acentuado o no.

  - Duración del alófono.

  - Índice de la palabra de la frase a la que pertenece el alófono.

* **Vector de F0:** representa la evolución temporal de la frecuencia fundamental, mediante una serie de valores en hertzios. Se obtienen los valores de F0 de manera síncrona: 1 valor cada 5 mseg.

Queremos resaltar que las distintas tareas del proceso lingüístico no siempre pueden completarse una a continuación de la otra, de manera secuencial. Muchas veces para poder completar una tarea será preciso disponer de alguna información que se generará más adelante, por lo que parte de la tarea deberá ser repetida después de que se disponga de dicha información. Esto será posible gracias a que todos los módulos del proceso lingüístico pueden acceder a las diferentes estructuras de datos en cualquier momento, bien para consultar determinada información, o para actualizarla.

A continuación, hablaremos de cada uno de los módulos que componen el proceso lingüístico del conversor texto-voz AMIGO, presentando las tareas que realizan.

3.1 Módulo Normalizador

Su tarea principal es detectar y reunir un conjunto de caracteres en el texto de entrada. Este conjunto de caracteres forma la unidad de trabajo que se tomará como común a todos los módulos.

La unidad de trabajo que se ha definido recibe el nombre de frase, si bien no es exactamente lo que se entiende generalmente por frase. La frase abarcará una serie de palabras, hasta que se encuentre algo que marque el fin de frase. Hemos decidido que el fin de frase viene marcado por:

* Cierre de interrogación.
* Cierre de admiración.
* Dos puntos.
* Puntos suspensivos.
* Punto de fin de frase.
* Dos caracteres nueva línea separados o no entre sí por uno o más blancos y/o tabulados.
* Fin del fichero de entrada.

Empleamos el término *palabras* al referirnos a las agrupaciones de caracteres de la frase separados entre sí por una sucesión de blancos, tabulados y nuevas líneas, aunque no siempre son palabras propiamente dichas (compuestas únicamente por signos alfabéticos), pues podemos encontramos con signos ortográficos, abreviaturas, ...

Como ya hemos dicho antes, la tarea principal del módulo normalizador es aislar una frase en el texto de entrada para poder pasarla al resto de los módulos como unidad de trabajo. Además realiza otra serie de tareas complementarias:

* Normalizar la escritura de las palabras detectadas, para simplificar comparaciones y consultas en el resto del módulo y en los módulos siguientes. La normalización supone varias cosas. Por un lado, se almacenan las distintas palabras constituyentes de la frase separando unas de otras por un único espacio en blanco. Por otro lado, se modifica la escritura de las palabras de la frase, pasando todos sus caracteres componentes a letras minúsculas. Además, también se modifica la representación de las secuencias de signos ortográficos, para los casos de signos incómodos de reconocer tal y como aparecen normalmente representados (puntos suspensivos) o que presentan alguna ambigüedad en su interpretación (las comillas y los guiones de comentario, que pueden ser de apertura o de cierre). Con el cambio de representación, se resuelve la ambigüedad.

* Obtener una serie de datos acerca de a qué tipo de palabra pertenece cada una de las palabras de la frase. Esto permitirá en otros módulos dar un tratamiento diferente a las palabras, de acuerdo al tipo al que pertenezcan (por ejemplo, palabra propiamente dicha, abreviatura, signo ortográfico, ...).

El normalizador comienza por reunir los caracteres que constituyen una frase. Por ello, su principal ocupación es decidir cuándo se produce cualquiera de las circunstancias que provocan un fin de frase, que ya se citaron más arriba. Esto no es una tarea trivial. Como ejemplo, baste considerar el caso del punto ortográfico:

Un punto ortográfico no es fin de frase si dicho punto está inmediatamente precedido de una secuencia de caracteres que forman una abreviatura válida (o el principio de una abreviatura válida de más de una palabra). Tampoco será fin de frase si va precedido por al menos un dígito, y seguido por otro, como puede aparecer en expresiones numéricas como *103.228*. Otro caso en que un punto ortográfico no es fin de frase es en las iniciales de los nombres, como *el señor J.L. Serrano*. Para tratar este caso, hemos decidido que un punto ortográfico precedido por una letra mayúscula y seguido por algo que no sea un signo ortográfico, no es un punto de fin de frase. En cualquier otro caso, el punto ortográfico se considera fin de frase. Lo que queda por aclarar en este caso, es si el punto forma parte de puntos suspensivos o no. Dada la muy alta variabilidad con que pueden venir representados los puntos suspensivos, hemos considerado un criterio bastante relajado para su identificación. Para nosotros, los puntos suspensivos pueden venir representados por una secuencia de tres o más puntos, separados entre sí o no por un número arbitrario de espacios en blanco y/o tabulados.

3.2 Módulo de preproceso

Aunque el módulo normalizador ya ha decidido y agrupado lo que es una frase, separado los elementos que la constituyen y realizado una primera normalización de su escritura, la variabilidad del texto de entrada es todavía demasiado grande. El resto de los módulos sólo van a poder manejar palabras y signos ortográficos, y hay que reducir a esta representación cualquier otra cosa (por ejemplo: abreviaturas, números, representaciones de fechas, representaciones horarias, ...).

Por tanto la principal tarea de este módulo es la de reducir la complejidad (variabilidad) del texto. Además se realizan otra serie de tareas:

* **Silabicación:** La información sobre la división en sílabas de la palabra es necesaria para poder determinar la acentuación fonética. Además influye en la decisión de los alófonos, controla las reglas de entonación y pausado, e indica cómo deben tratarse las palabras ilegibles (siglas o palabras no válidas en castellano). La Real Academia proporciona unas reglas sobre cómo se debe realizar esta tarea, aunque algunos casos no quedan totalmente determinados. El conversor AMIGO emplea estas reglas, tal como aparecen descritas en [1], aunque hemos tenido que completarlas con otras reglas derivadas por nosotros. Si bien el conjunto original de reglas es suficiente para tratar correctamente las palabras del castellano, nuestro interés incluía también que se tratasen de una manera razonable las palabras que no son propias del castellano, como siglas y palabras extranjeras, y que se facilitase el tratamiento de palabras impronunciables.

* **Acentuación fonética:** En castellano la mayor parte de las palabras presentan un único acento fonético, cuya posición queda perfectamente determinada por la grafía de la palabra y su división en sílabas.

  Normalmente, una palabra tendrá un acento, pero no siempre es así.

  Hay algunas palabras que presentan dos acentos. Éstas son los adverbios procedentes de adjetivos, terminados en *mente*. Un acento recae sobre la raíz de la palabra que corresponde al adjetivo, y el otro en la primera *e* de *mente*.

También hay palabras inacentuadas. Suelen ser palabras función (preposiciones, artículos determinados, etc), sin contenido semántico por sí mismas. Sin embargo, hay palabras con la misma grafía que estas palabras función, pero que son verbos, nombres, etc, y que por tanto sí deben ser acentuadas. Este módulo lo único que puede hacer es marcarlas como ambiguas, y hay que esperar a que se haya realizado el análisis de categorías para decidir la acentuación correcta.

Hay otras particularidades de la acentuación que se tendrán en cuenta en otros módulos, como es la supresión del acento en algunas de las palabras componentes de una cantidad numérica, o el impedir que se dé una palabra átona antes de pausa.

*__Tratamiento de acrónimos y secuencias impronunciables:__ Cuando se va a leer una secuencia de letras sin las suficientes vocales para pronunciarla fácilmente, normalmente se suele deletrear (al menos la parte impronunciable de la secuencia). Por ejemplo: *ftp* se lee *efe te pe*, *HB* se lee *hache be*, *frnos* seguramente se leería como *efe erre nos*.

En los casos en que sea necesario, se hace una segunda silabicación más ajustada a esta tarea, y luego se deletrean las sílabas que quedan sin vocal. Además se tiene en cuenta la forma en que venía escrita la palabra, algunas reglas heurísticas que hemos generado, y un conjunto de excepciones que el usuario puede adaptar a sus necesidades.

3.3 Módulo Categorizador

La principal tarea del módulo es la de asignar categorías a las palabras. Las categorías que se asignan no son exactamente categorías gramaticales, sino un conjunto de códigos que en muchos casos se corresponden con verdaderas categorías gramaticales, pero que en otros se han escogidos por criterios prácticos.

Hay algunas partes de la categorización que ya se han realizado en el módulo de preproceso, como es el caso de las palabras con la inicial en mayúsculas no siendo principio de frase (nombres propios), los resultados de las expansiones (números, letras, signos) y las palabras terminadas en *mente* (el acentuador fonético debe saber si son adverbios o no, para decidir el número correcto de acentos).

La última versión del módulo categorizador presenta numerosas mejoras respecto a la versión anterior (descrita en [2]). La principal de ellas consiste en la renuncia a dejar a la salida del módulo categorías consideradas como ambiguas. Cuando en la anterior versión no se podía categorizar una palabra con una cierta seguridad, se dejaba como ambigua dentro de un grupo de categorías que reflejaban esas ambigüedades. De esta manera, aunque el funcionamiento del módulo era correcto, los resultados no eran muy útiles para otras tareas (pausado), salvo algunos tipos de ambigüedades muy concretos.

La tarea de categorización se puede descomponer en tres tipos de subtareas: rutinas que buscan en tablas la palabra a categorizar (menos de 3500 palabras recogidas en las tablas), rutinas que comprueban la estructura de la palabra para intentar descubrir su categoría, y rutinas que comprueban la relación de la palabra en cuestión con otras palabras de su entorno.

Aunque estos tipos de rutinas ya se empleaban en la anterior versión, se han diversificado, se ha aumentado el rigor y se han aprovechado mejor sus posibilidades. En esta mejora general se puede destacar una mejor utilización de las locuciones, un aumento considerable en el número y calidad de las reglas de contexto, una mejor estructuración de las tareas en el tiempo, y sobre todo,un análisis mucho más profundo y detallado de las formas verbales.

Como muestra de las mejoras del nuevo módulo de pausado, presentamos una comparación de los resultados obtenidos con la antigua versión y con la nueva.

	Versión antigua	Versión nueva
Palabras correctamente etiquetadas	78%	96.9%
Palabras incorrectamente etiquetadas	4%	3.1%
Palabras con categoría ambigua	18%	0.0%

Los resultados se obtuvieron contrastando los resultados dados por el categorizador con los asignados manualmente por 3 personas (3 de los autores del presente artículo), sobre textos que suman 10168 palabras.

3.4 Módulo Estructurador

La misión de este módulo es extraer información sobre la estructura sintáctica de la frase. Este módulo ha sido terminado de desarrollar muy recientemente, y todavía no se han modificado los módulos siguientes para aprovechar la información sintáctica generada por él.

En la comunicación *Módulo de análisis sintáctico para un sistema de conversión texto-voz en castellano*, presentada en este mismo congreso, se realiza una descripción del módulo estructurador del conversor texto-voz AMIGO.

3.5 Módulo Pausador

Al leer un texto, es muy frecuente que los lectores hagan más pausas que las que vienen marcadas por los signos ortográficos. Este hecho normalmente viene motivado por la necesidad que tiene el hablante de recuperar el aliento.

Evidentemente, un conversor texto-voz no necesita recuperar el aliento cuando procesa un texto, pero daría una desagradable sensación de ahogo al oyente si pronunciara un fragmento de texto largo que no tuviera ningún signo ortográfico entre medias. Hemos podido comprobar que fragmentos de texto de este tipo aparecen frecuentemente en noticias de periódicos, artículos de opinión, novelas, ... por lo que no puede decirse que siempre sean producto de un incorrecto (o al menos escaso) uso de los signos de puntuación.

Por otro lado, la conversión de grafemas a alófonos se hace mediante unas reglas, algunas de las cuales consultan la existencia de pausa delante o detrás de la letra bajo estudio. Así, la conversión grafema-alófono depende en algunos casos de las pausas, sean estas marcadas ortográficamente o no.

Son bastantes los autores ([3], [4], [5]) que aceptan que la prosodia de un texto se ve influida en cierta manera por la estructura sintáctica de ese texto. Nosotros participamos de esa opinión, y por ello nos decidimos a realizar el módulo estructurador que fue comentado anteriormente. Pero todavía no hemos desarrollado un nuevo pausador que considere la información sintáctica, tarea que abordaremos en breve.

El pausador que ahora tenemos no llega a hacer un análisis sintáctico del texto, pero sí realiza una división del texto en una serie de fragmentos que, normalmente, tienen una cierta unidad sintáctica. De esta manera, el pausado ayuda a la generación de una prosodia más correcta y así se puede dar la impresión de que el conversor *entiende* lo que está leyendo, aumentando la naturalidad de la voz sintética recibida por el oyente.

La tarea principal del módulo pausador, como hemos dicho antes, es insertar pausas automáticamente cuando el texto que se encuentra entre dos pausas marcadas ortográficamente produciría una secuencia de habla demasiado larga. Además realiza otras tareas complementarias:

*Identificar las secuencias de signos ortográficos dentro de una frase, y asignarles un código único a cada una. Ese código único es el que caracterizará la pausa que cada secuencia de signos ortográficos debe introducir en el texto.

*Corregir la acentuación fonética de las palabras de acuerdo a criterios prosódicos. Hemos podido detectar que las palabras átonas que se encuentran delante de una pausa no se pronuncian realmente como átonas. Por ello, siempre que se encuentre una pausa (ortográfica o no) se comprueba si la palabra anterior es átona, en cuyo caso se acentuará fonéticamente y se marcará como tónica.

Actualmente se consideran 24 códigos de pausa diferentes: 16 correspondientes a pausas generadas por signos ortográficos, 7 correspondientes a pausas no generadas por signos ortográficos, y un código adicional para indicar ausencia de pausa.

Una vez procesadas las pausas ortográficas, la frase puede quedar dividida en varios trozos, limitados por dichas pausas.

Para cada uno de los trozos (al menos habrá uno) se calculan el número de sílabas y el número de sílabas tónicas, y se comprueba que ambas cantidades estén por debajo de unos límites prefijados. Después de varias tentativas, hemos obtenido unos buenos resultados fijando esos límites en 22 sílabas y 8 tónicas.

En caso de que el trozo tenga 22 o más sílabas, u 8 o más tónicas, se intenta introducir una pausa. La búsqueda para la inserción de la pausa se hace entre la palabra inicial del trozo y la que está separada de ella al menos 22 sílabas u 8 tónicas.

Dentro del intervalo de búsqueda, se intentan localizar palabras con unas categorías determinadas que pueden inducir la inserción de una pausa. Esas categorías son las siguientes, ordenadas de mayor a menor prioridad:

*Conjunciones coordinantes.
*Conjunciones subordinantes.
*Verbos.
*Palabras función.

Cada una de esas categorías contemplan una serie de excepciones que impide la introducción de pausa. Sólo en caso de que no se haya podido introducir pausa para una categoría, se pasa a la siguiente categoría de prioridad más baja.

Como hemos dicho antes, actualmente el pausador siempre genera pausa en las posiciones de los signos ortográficos. Sin embargo, hemos podido comprobar que, en ocasiones, un lector humano puede no hacer pausa al encontrar algunos signos ortográficos y la lectura sigue resultando perfectamente correcta y natural. Convendría estudiar este fenómeno para saber si es posible reflejarlo en el pausador.

3.6 Módulo Conversor de Grafemas a Alófonos

Este módulo se ocupa de obtener la secuencia de alófonos correspondiente a la secuencia de letras de una frase dada. Para ello empleará parte de la información lingüística obtenida en los módulos anteriores.

El conjunto de alófonos que hemos considerado es el que aparece en [6], y de ahí hemos tomado también los criterios para construir nuestras reglas de conversión.

Básicamente, la información empleada en las reglas de conversión es la siguiente:

*Caracteres (letras) componentes de la frase.
*Límites entre palabras dentro de la frase.
*Límites entre sílabas dentro de las palabras.
*Acentuación fonética de las palabras.
*Localización de las pausas.

Con nuestras reglas de conversión se obtienen transcripciones fonéticas *académicamente correctas*.

Posteriormente, se hace una transformación de los alófonos obtenidos para pasar a una versión de la transcripción fonética en un alfabeto limitado a 26 alófonos, frente a los 45 del alfabeto original. Se ha hecho así por consideraciones prácticas de diseño del conjunto de unidades empleado en el bloque de síntesis de voz, que se explicarán más adelante.

3.7 Módulo Generador de los Parámetros Prosódicos

La tarea que se realiza en este módulo es la de generar los parámetros que van a determinar la prosodia. Esta tarea se puede descomponer en otras dos, una por cada parámetro: duración y entonación (contorno de frecuencia fundamental).

Se asignan valores de duración a las pausas (tanto las ortográficas como las introducidas por el módulo de pausado) y a los alófonos.

El modelo de duración de las pausas trata 18 tipos diferentes (cualquier otro tipo de pausa tiene que ser reducido a uno de estos 18). Dependiendo únicamente del tipo de pausa, se asigna un valor medio y un margen, a partir del cual se genera una desviación aleatoria con una distribución uniforme que se suma al valor medio para dar la duración definitiva. Este modelo procede del descrito en [7], que hemos refinado un poco con las mismas ideas básicas.

Para asignar duraciones a los alófonos disponemos de un modelo más complejo. Se hace depender la duración de cada unidad de 5 factores: el alófono considerado, los alófonos adyacentes, la posición respecto al acento, la posición del alófono dentro del grupo fónico al que pertenece, y la longitud del grupo fónico. Cada alófono cuenta con una duración base propia que es multiplicada por cantidades derivadas de los 5 factores anteriores. La duración resultante puede ser superior o inferior a la de base. Este modelo aparece descrito en [8].

La entonación es el otro parámetro considerado en la generación de la prosodia, y es el que más influye en la calidad y naturalidad de la voz producida. Entonación es un término relacionado principalmente con la fonología y la percepción, mientras que este módulo únicamente va a generar valores de la frecuencia fundamental o F0 sin considerar efectos de un nivel superior. Sin embargo, usaremos también el término entonación o tono para referirnos a este parámetro, pues es muy habitual hacerlo así.

El bloque de proceso de señal necesita recibir información sobre el tono con un muestreo uniforme: un valor por cada 5 milisegundos de habla. Para generar estos valores se utiliza un modelo que asigna valores de F0 a determinados puntos de las sílabas (normalmente a los puntos medios de los núcleos de las sílabas tónicas y algunas átonas), y luego se interpola de acuerdo a los valores de duración, para generar las muestras necesarias. Este modelo procede del descrito en [9], si bien ha sido bastante modificado, corregido y ampliado. A pesar de esto, los principios en los que se basa siguen siendo los mismos.

4 BLOQUE DE SÍNTESIS DE VOZ

El sintetizador que empleamos es de los llamados de concatenación de unidades. Estos sintetizadores se basan en tener un conjunto de pequeños segmentos de voz tomados de un hablante, que se van concatenando para formar el discurso deseado. Naturalmente, la concatenación debe ser

controlada, de manera que, por un lado, se eviten discontinuidades y sonidos anómalos y, por otro lado, se puedan variar los elementos prosódicos respecto a aquellos que originalmente tenían los segmentos de voz escogidos. Para decidir el tamaño y número de estas unidades hay un compromiso entre la calidad de la voz que se quiere sintetizar, y limitaciones de memoria de datos.

La posibilidad más inmediata es tener grabados únicamente cada uno de los alófonos. Así sólo se necesitaría tener grabadas entre 20 y 50 unidades, según el alfabeto de alófonos considerado. Por contra, habría que realizar el *pegado* de las unidades en zonas acústicamente inestables, como son las transiciones entre alófonos. Por esto, la aproximación más común es trabajar con demialófonos, que son unidades que recogen parte de un alófono, la zona de transición, y parte del siguiente alófono. Así, el *pegado* se hace en zonas acústicamente más estables, como son las partes centrales de los alófonos.

Como no siempre los alófonos llegan a alcanzar zonas acústicamente estables, extendiéndose los fenómenos de coarticulación de un alófono más allá de los alófonos inmediatamente adyacentes, nuestra colección de unidades para la síntesis contempla diferentes tipos de elementos: trozos de voz de longitud inferior a un alófono completo (sub-alófonos), demialófonos, trialófonos y tetraalófonos.

En el apartado anterior dijimos que generábamos un alfabeto de alófonos reducido a 26 elementos por motivos prácticos de diseño del inventario. Cuantos más alófonos distintos se consideren, aumentará mucho más el número de sus posibles combinaciones y, por tanto, el número de unidades del inventario. Además, hemos encontrado alófonos que realmente no eran sino variantes de otros debidas a su contexto más cercano, y ese efecto queda en muchas ocasiones recogido en las unidades del inventario.

Actualmente contamos con un inventario de 725 unidades, con el que hemos conseguido una calidad acústica bastante satisfactoria. El diseño de una colección de unidades no es, de ninguna manera, una tarea inmediata, y requiere sucesivas revisiones y ajustes para mejorar la calidad de la voz sintética, muchas veces de una forma artesanal más que teórica, si bien hay ciertos criterios que conviene tener en consideración (como los expuestos en [10]).

La algorítmica del bloque de síntesis de voz de nuestro conversor texto-voz AMIGO es la misma que se emplea en el conversor texto-voz en inglés desarrollado previamente por los laboratorios Bell de AT&T, con los que hemos estado mantenido una colaboración prolongada. El tipo de codificación empleado para las unidades es predicción lineal multipulso (MPLPC) con 16 coeficientes.

El procedimiento de síntesis comprende tres etapas diferenciadas:

* Escoger la secuencia de unidades del inventario que se corresponde con la secuencia de alófonos que se quiere sintetizar. Dada la composición del inventario y los distintos tipos de unidades que en él hay, puede haber distintas secuencias posibles de unidades del inventario para una determinada secuencia de alófonos. El criterio que se emplea para seleccionar la secuencia de unidades del inventario es buscar aquella en la que los alófonos se encuentren, siempre que sea posible, completos dentro de una unidad; es decir, siempre se prefieren las unidades de mayor longitud frente a las de menor longitud (tetraalófonos frente a trialófonos, trialófonos frente a demialófonos, ...) en caso de que se pueda elegir. Por ejemplo, para la secuencia de alófonos *vwxyz*, se prefiere la secuencia de unidades *vwxy-yz* frente a la secuencia de unidades *vwx-xy-yz*, en caso de que ambas sean posibles.

* Adecuar el número de tramas de cada alófono a la duración que le haya sido asignada. Para esto habrá que tener en cuenta las tramas del alófono correspondientes a dos unidades diferentes, cuando el alófono no se encuentre en la posición intermedia de una unidad. Es decir, en el caso de la secuencia de unidades *vwxy-yz*, las tramas componentes del alófono *y* se encuentran en parte en la unidad *vwxy* (transición de *x* a *y*, y primera parte de *y*) y en parte en la unidad *yz* (segunda parte de *y*, y transición de *y* a *z*). Si la duración asignada es mayor que la duración derivada de las unidades del inventario, se generan nuevas tramas mediante interpolación de las existentes. Si la duración asignada es menor que la duración derivada de

las unidades del inventario, se eliminan algunas tramas. La interpolación y la eliminación de tramas debe hacerse de forma selectiva, especialmente la eliminación, pues de no hacerse así se podrían perder características esenciales de algunos sonidos, como es el caso de los primeros instantes de la explosión en las oclusivas sordas.

* Producción de las muestras de voz. Una vez ajustado el número de tramas de síntesis para todos los alófonos de la frase, se pasan las mismas al algoritmo que produce las muestras de voz, pero imponiendo los valores de F0 calculados en el módulo generador de los parámetros prosódicos. La síntesis se realiza mediante técnicas estándar de predicción lineal multipulso. Para las tramas sonoras de dispone de una excitación multipulso prototípica, que es modificada convenientemente de acuerdo a los parámetros prosódicos. Para las tramas sordas, se genera una excitación multipulso aleatoria.

Para una descripción más completa del inventario de unidades, su estructura y otros aspectos del bloque de síntesis de señal, se puede consultar [11].

5 RESULTADOS Y TRABAJOS FUTUROS

Hay partes del conversor AMIGO para las que pueden hacerse medidas más o menos objetivas de su calidad, como detectar bien las frases, hacer una correcta división en sílabas de las palabras, realizar un buen etiquetado gramatical de las palabras, ...

Pero esta evaluación de tipo objetivo, si bien es necesaria, no dice mucho acerca de la calidad global del conversor. Las medidas más adecuadas para enjuiciar esta calidad necesariamente pasan por comparar la calidad de la voz sintética producida por el conversor con la de voz natural producida por un hablante humano. Este tipo de medidas se hacen normalmente con criterios subjetivos, en el sentido de que lo más importante es la opinión que acerca de la calidad del conversor tenga un conjunto suficientemente amplio de oyentes humanos.

Existen una serie de experimentos más o menos normalizados que intentan evaluar tanto la inteligibilidad como la naturalidad de este tipo de sistemas, aunque estos experimentos han sido desarrollados para otros idiomas y pueden ser más a menos favorables según el tipo de conversor texto-voz que se considere.

Actualmente no disponemos de un estudio formal acerca de la calidad del conversor AMIGO, y es algo que consideramos muy importante para futuras mejoras.

Hasta el momento sólo se han hecho unas 500 encuestas a los usuarios experimentales de un servicio de noticias a través del teléfono (Audiotex), en el que la información se daba mediante el conversor texto-voz AMIGO. En esas encuestas se pidió que los usuarios enjuiciasen la inteligibilidad y la naturalidad del conversor, en una escala de 1 a 5, en la que 1 era la peor puntuación y 5 la mejor. Los resultados obtenidos fueron de un 4,5 en inteligibilidad y un 3,1 en naturalidad. Aunque estos resultados no tienen mucho valor en sí mismos por la informalidad del estudio, los consideramos muy alentadores.

Todavía quedan abiertas bastantes vías de mejoras para el conversor AMIGO, especialmente en lo que concierne a la obtención de una prosodia más natural. Pensamos que a este respecto puede ser bastante importante la introducción de más información sintáctica tanto en el pausado automático como en los modelos de duraciones de los alófonos y de generación de los contornos de F0.

También debemos seguir mejorando la calidad del inventario de unidades, aumentando el número de las mismas y optimizando los criterios para su selección.

Por último, aunque la calidad acústica obtenida con el modelo de síntesis actual es bastante satisfactoria, se pueden buscar otros modelos que o bien sean más flexibles en cuanto a sus posibilidades de control de la voz sintética, o bien estén más adaptados a las particularidades de otros tipos de voz (voz femenina, por ejemplo).

6 REFERENCIAS

[1]J.A. Mañas, *Word division in Spanish*, Communications of the ACM, vol 30, nº 7 (1987).

[2]M.Á. Rodríguez, J.G. Escalada, *Text analysis system with automatic letter to allophone conversion for a Spanish text-to-speech synthesizer*, proceedings of the ESCA workshop on speech synthesis (1990).

[3]J. 't Hart, R. Collier, *Integrating different levels of intonation analysis*, Journal of Phonetics 3 (1975).

[4]G. Bailly, *Integration of rhythmic and syntactic constraints in a model of generation of French prosody*, Speech Communication, 8 (1989).

[5]D.H. Klatt, *Vowel lengthening is syntactically determined in a connected discourse*, Journal of Phonetics 3 (1975).

[6]A. Quilis, *El comentario fonológico y fonético de textos*, Arco/libros (1988).

[7]A. Santos y otros, *Diseño y evaluación de reglas de duración en la conversión de texto a voz*, boletín nº 6 SEPLN (1988).

[8]A. Macarrón y otros, *Generation of duration rules for a Spanish text-to-speech synthesizer*, proceedings of Eurospeech 91 (1991).

[9]P.J. Moreno y otros, *Improving naturalness in a text-to-speech system with a new fundamental frequency algorithm*, proceedings of Eurospeech 89 (1989).

[10]A. Macarrón, *Design and generation of the acoustic database of a text-to-speech synthesizer for Spanish*, proceedings of the ESCA workshop on speech synthesis (1990).

[11]J.P. OLive, *A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds*, proceedings of the ESCA workshop on speech synthesis (1990).