VOLUMEN MONOGRÁFICO (1999), 103-116

TECNOLOGÍAS DEL HABLA Y SÍNTESIS DE VOZ EN GALLEGO

ELISA FERNÁNDEZ REI
Centro Ramón Piñeiro para la Investigación en Humanidades

RESUMEN. *Este capítulo se divide en dos grandes partes: la primera, de carácter introductorio, sitúa las Tecnologías del Habla dentro de las Industrias de la Lengua y se hace mención de la importancia que éstas tienen para la normalización del gallego; también se dedica un pequeño apartado a describir brevemente algunos productos de las Tecnologías del Habla, como el reconocimiento de voz y los sistemas de gestión de diálogo. La segunda parte, que constituye el grueso del capítulo, está dedicada a la Síntesis de Voz: tanto desde un punto de vista más o menos general (aplicaciones de este sistema y tipos de sintetizadores existentes), como desde una perspectiva más particular, deteniéndose en la descripción del conversor texto-habla en gallego, cuya arquitectura se detalla con toda la exhaustividad que permite el espacio del que disponemos.*

PALABRAS CLAVE. *Tecnología de la voz, industrias de la lengua, fonética acústica.*

ABSTRACT. *The chapter is divided in two parts. The first, introductory in nature, defines speech technology as part of computational linguistics and highlights the importance of this technology to the development of Galego, an endangered language. This introduction also touches upon speech technology applications such as speech recognition and dialogue management systems. The second part, which forms the larger portion of the chapter, is dedicated to speech synthesis. This part begins with a description of existing speech synthesis applications and types of synthesizers. It goes on describe the architecture of the text-to-speech conversion system (TTS) in Galego in as much detail as the space allotted to the chapter permitted.*

KEYWORDS. *Speech technology, language engineering, acoustic phonetics.*

1. INTRODUCCIÓN

A medida que el lector avance en la lectura de este capítulo, se dará cuenta de que el tema central no es tanto las Tecnologías del Habla en general, como la descripción de un sintetizador de voz en particular. La razón de detenernos con más detalle en la conversión texto-habla es que es ésta la especialidad más desarrollada en

gallego en la actualidad, dentro del ámbito de las tecnologías del habla, por lo que nos pareció más interesante centrarnos en un producto concreto que, además, tenemos la suerte de conocer de cerca.

Con todo, sería conveniente situar previamente las Tecnologías del Habla dentro del campo más amplio de la Lingüística Computacional. Esta disciplina o conjunto de disciplinas, que también recibe otras denominaciones como Ingeniería Lingüística, Industrias de la Lengua, Lingüística Informática o Tecnología Lingüística, conjuga los conocimientos y métodos de la Lingüística, por un parte, y de la Tecnología de la Información, por la otra. De este modo, la relación entre Lengua y Tecnología genera un importante número de especialidades, productos y aplicaciones: correctores ortográficos y gramaticales para procesadores de texto, diccionarios electrónicos para diversos usos, traducción automática, síntesis y reconocimiento del habla y un largo etcétera.

Sin embargo, esta relación que se establece entre lengua y tecnología es más restringida en el campo de las Tecnologías del Habla, en tanto que de la lengua sólo le interesa la vertiente oral, el habla. Por tanto, la Tecnología del Habla surge del interés en hacer del habla, de la voz, un medio de comunicación efectivo no sólo entre las personas, sino también entre el ser humano y la máquina. Podríamos definirla, pues, como un conjunto de conocimientos y técnicas de procesamiento de la señal de voz dirigidos a aumentar su capacidad de comunicación.

Así, el objetivo último de las Tecnologías del Habla es proporcionar los medios, técnicas y procedimientos necesarios para que la comunicación oral entre personas y máquinas sea lo más semejante posible a la que se da entre las personas. Este objetivo ideal presenta una serie de ventajas, entre las que podríamos destacar, la libertad de movimientos que supone una comunicación con las máquinas exclusivamente oral o la facilidad de acceso a ellas, incluso por vía telefónica. Sin embargo, aún hay numerosos inconvenientes, como la escasa naturalidad del habla que utilizan estos sistemas hasta el momento o las importantes limitaciones de algunos de sus productos, como por ejemplo el reconocimiento del habla continua.

La combinación de distintas tecnologías desarrolladas en el ámbito de las Industrias de la Lengua abre unas perspectivas amplísimas para el futuro: eliminaría las barreras lingüísticas entre personas y entre personas y máquinas, y permitiría establecer una nueva relación con los ordenadores (tanto de uso público como PCs), en la medida en que no se necesitaría estar físicamente delante de ellos para acceder a los servicios que proporcionan o a la información que contienen, ya que nos podríamos establecer comunicación a través del teléfono. Así, la traducción automática, por ejemplo, no tiene por qué limitarse exclusivamente al texto escrito, pues combinada con la síntesis y el reconocimiento daría lugar a auténticos intérpretes automáticos.

No obstante, antes de pasar a analizar con detenimiento algunas de estas técnicas, convendría hacer una aclaración: el estado actual de la investigación no se

corresponde con la definición ideal que demos de ninguna de estas aplicaciones: esto es, ni la síntesis actual es capaz de producir textos comparables en calidad a los producidos por un lector natural, ni el reconocimiento de voz consigue escribir lo que "oye" como una persona, ni tampoco los programas de traducción pueden traducir un texto cualquiera de una lengua a otra con la calidad que lo hace un traductor humano. No son, pues, descripciones del estado actual de la investigación, sino definiciones que se corresponden con el objetivo último o deseable de la investigación en los distintos campos.

Por último, por lo que se refiere a las aplicaciones de la Tecnología Lingüística, cabe mencionar que, además de las comerciales, terapéuticas, médicas o didácticas, hay una serie de aplicaciones propiamente lingüísticas, ya que la programación y el diseño lingüístico de sistemas informáticos tiene como finalidad no sólo dotar a la máquina de competencia lingüística, sino también dotar a la lingüística de medios computacionales. De este modo, por un lado, la Lingüística Teórica se ve impulsada por las necesidades planteadas por las Industrias de la Lengua y, por otro lado, los productos de las nuevas tecnologías son, en muchas ocasiones, instrumentos muy útiles para la validación de algunas teorías formuladas por la Lingüística Teórica.

2. LA IMPORTANCIA DE LAS INDUSTRIAS DE LA LENGUA PARA LA NORMALIZACIÓN DEL GALLEGO

Todo parece apuntar a que, ya en el presente y en mayor medida en un futuro próximo, las telecomunicaciones van a modificar notablemente el mercado, el marketing y el comercio en general. Por consiguiente, dado que las nuevas tecnologías aplicadas a las telecomunicaciones utilizan necesariamente una lengua como soporte, la presencia de las lenguas en este ámbito va a ser determinante para la consecución y/o conservación de un status de lengua normal o normalizada. Con todo, la preocupación por estar presentes en las aplicaciones de las nuevas tecnologías es común a todas las lenguas, no sólo a las minorizadas o a las que cuentan con un pequeño número de hablantes, sino también a las de reconocido prestigio y expansión internacional, como el francés, el español o el alemán.

En las lenguas minorizadas, como el gallego, la consecución de un status de lengua normal pasa por la ampliación de sus usos hasta conseguir su presencia en todos los ámbitos. Así pues, es especialmente relevante en su caso la incorporación de los desarrollos derivados de las Tecnologías del Habla, y de las Industrias de la Lengua en general, ya que esto supondría la presencia del gallego en las múltiples aplicaciones de estas tecnologías con las consiguientes elevación de su prestigio y extensión de usos. En el caso contrario, la ausencia del gallego en estos nuevos ámbitos de uso no sólo pondría en peligro su normalización, sino que, de verse

definitivamente alejada de los avances técnicos que se están produciendo en la actualidad, podría ver amenazada su propia supervivencia.

3. TECNOLOGÍAS DEL HABLA

La multiplicidad y variedad de especialidades derivadas de las Tecnologías del Habla traen como consecuencia la necesidad de conjugar saberes muy distintos para desarrollar los distintos productos y aplicaciones, de ahí la importancia que adquiere la interdisciplinaridad y el trabajo conjunto que llevan a cabo ingenieros, médicos, lingüistas, psicólogos...

Hay múltiples áreas de trabajo en Tecnología del Habla, desde la síntesis de voz o el reconocimiento de habla, hasta la codificación de voz u otras técnicas relacionadas, como la reducción de ruidos, análisis de voz con fines médicos y patológicos, desarrollo de prótesis auditivas para ayudar en la locución, etc.

La codificación de voz abarca desde las técnicas más sencillas que permiten el almacenado y acceso de los ordenares a la señal de voz, hasta las más complejas que no sólo permiten un importante ahorro de recursos, sino que también ofrecen representaciones y modelos útiles a todas las Tecnologías del Habla.

La síntesis de voz, de la que se hablará con más detalle en el apartado siguiente, es la conversión de un texto escrito en una cadena oral de un modo totalmente automatizado, sin la intervención del hablante.

El reconocimiento de voz es el proceso inverso al de la síntesis, pues convierte, también mediante procesos totalmente automatizados, la señal de voz en texto o en algún tipo de mensaje o código manejable por el ordenador.

Según el estilo de habla que se vaya a reconocer, se distinguen tres tipos de sistemas:

- reconocimiento de palabras aisladas: el usuario debe pronunciar una sola palabra o comando, precedidos y seguidos de un silencio, de entre los recogidos en la aplicación.
- reconocimiento de palabras conectadas: el usuario pronuncia de forma fluida utilizando un vocabulario restringido.
- reconocimiento de habla continua: el usuario pronuncia frases de manera natural, con un vocabulario muy amplio.

Otro de los factores que definen el sistema de reconocimiento es la dependencia que éste tenga del locutor: hay sistemas que incorporan patrones adaptados a un locutor determinado y, por tanto, sólo funcionan correctamente para él y hay otros sistemas en los que los patrones pretenden ser válidos para cualquier hablante. Los primeros requieren un entrenamiento del sistema con una cantidad de voz reducida, pero suficiente, que permita adaptarlos a un hablante particular.

Por último, el tamaño y la dificultad de la tarea es otro factor que va a condicionar también la estructura del sistema, ya que el número de palabras recogidas en el vocabulario de la aplicación puede variar mucho y, además, el grado de similitud fonética entre estas palabras puede complicar bastante la tarea.

En cuanto a los sistemas de gestión de diálogo, son un conjunto de módulos (reconocimiento del habla, analizador, base de datos, generador de unidades lingüísticas, síntesis del habla), ideados para realizar, cada uno de ellos, una tarea concreta. Están conectados a un módulo principal que lleva a cabo el control, o gestión propiamente dicha, del diálogo: mantiene la coherencia entre la pregunta del usuario y el sistema, resuelve las anáforas y elipsis que se producen en el discurso normal, genera respuestas, intenta predecir, dentro de lo posible, las reacciones del usuario...

4. LA SÍNTESIS DE VOZ

Como ya señalábamos más arriba, un sintetizador de voz es una herramienta que permite convertir un texto escrito en una cadena oral sin la intervención directa del hablante.

Los conversores están sometidos a una serie de condicionantes tanto lingüísticos como tecnológicos y uno de los objetivos es obtener un resultado que sea fruto de un buen compromiso entre estos dos tipos de condicionantes. Así, por ejemplo, la relación establecida entre complejidad y rapidez debe ser óptima y, en gran medida, ventajosa para la segunda, para que su funcionamiento se dé en tiempo real.

Generalmente, se utilizan una serie de parámetros para la caracterización y evaluación de los sistemas de síntesis: la naturalidad, la calidad y inteligibilidad del habla sintetizada y la versatilidad del sistema. No debemos olvidar que el resultado de la conversión debe ser no sólo inteligible, sino también lo más natural posible. Asimismo, el sistema tiene que ser sucifientemente flexible para que sea capaz de soportar el máximo número de aplicaciones posibles.

4.1. *Aplicaciones*

Como ya avanzaba en el apartado anterior, existe una relación obvia entre el sistema y la aplicación. Puede construirse un sistema más o menos dependiente de la aplicación a la que se vaya a dedicar y, por lo tanto, más o menos versátil y más o menos eficaz para cada aplicación concreta. Así pues, limitando el sistema a una aplicación particular, quizás se consiga una menor complejidad, pero también poseerá menos flexibilidad. Si por el contrario, el mismo sistema se puede utilizar para realizar tareas de distinta índole, ganará en versatilidad, pero probablemente será menos efectivo. Una vez más debemos intentar conseguir un buen compromiso entre flexibilidad, versatilidad y eficacia.

Las aplicaciones de la síntesis de voz no sólo se limitan al campo de la lingüística (validación de teorías o hipótesis fonológicas, por ejemplo), sino que sus aplicaciones suponen en muchos casos una verdadera revolución en el mundo de la comunicación e incluso en el mundo de las relaciones sociales.

Podríamos dividir las aplicaciones de los conversores en dos grandes grupos, según los objetivos que cumple la síntesis de voz en uno u otro caso:

a) síntesis de voz para puestos de información telefónica automatizada (información meteorológica, administrativa, académica, turística...): permite la actualización constante de la información y, al mismo tiempo, abarata los costes, en tanto que este sistema permite tener varias líneas telefónicas para una misma información.

b) síntesis de voz para la ayuda a discapacitados vocales y auditivos: concede a los invidentes la posibilidad de acceder a la información textual residente en ordenador y a las personas incapacitadas para hablar les permite sintetizar voz utilizando un teclado, que puede incluso ser un teclado especial cuando la discapacidad afecta también a las facultades motrices. En ocasiones, se ha utilizado con finalidades terapéuticas (corrección de la dislexia) e incluso didácticas (por ejemplo, en inglés ha sido utilizado para la enseñanza de ortografía y otras asignaturas como álgebra).

La síntesis de voz ha alcanzado el grado de madurez necesario para soportar gran variedad de aplicaciones, pese a que aún posee limitaciones. Ya mencionamos antes que los sintetizadores actuales no igualan la calidad de la voz natural, pues no tienen, entre otras cosas, la variedad de registros o de matices de entonación que posee un hablante natural, por lo que sus aplicaciones son por el momento limitadas. De este modo, el gran reto de los sintetizadores es alcanzar una calidad óptima y un grado de naturalidad cercano al del habla humana, para que sean capaces de proporcionar un servicio de calidad, cómodo y eficaz.

4.2. *Tipos de sintetizadores*

Generalmente, la clasificación de los sintetizadores se lleva a cabo según el método de generación de la voz sintética. Existen en la actualidad tres grandes grupos de conversores texto-habla:

a) de formantes: consisten en una composición de filtros que modelan las resonancias y antirresonancias de las cavidades vocal y nasal. Son muy flexibles y producen voz sintética de gran calidad, si se realiza un

ajuste manual de los parámetros; por el contrario, para la síntesis automática, necesitan un número de reglas excesivo.

b) articulatorios: intentan simular la propagación de las ondas acústicas en el tracto vocal. Este modelo tiene la desventaja de que también es bastante complejo y, además, se carece de los conocimientos de fonética articulatoria necesarios para desarrollar con éxito este tipo de sintetizadores.

c) de concatenación de unidades: son los más usados hoy en día, debido a su buen compromiso entre complejidad y prestaciones. Este método consiste en almacenar segmentos de voz natural que, posteriormente, se unen (concatenan) para formar la cadena hablada. Necesitan un algoritmo que permita, además de la concatenación de las unidades, modificar prosódicamente los segmentos que se van a concatenar.

4.3. *El conversor texto-habla gallego*

El sintetizador de voz que se está elaborando en el Centro Ramón Piñeiro para la Investigación en Humanidades de Santiago de Compostela es fruto de un proyecto que tiene como característica primordial la interdisciplinaridad, ya que el equipo investigador que lo está confeccionando lo integran lingüistas de la Universidade de Santiago (Manuel González González, Rut Losada Soto y Elisa Fernández Rei) e ingenieros de la Escola Técnica Superior de Enxeñeiros de Telecomunicacións de la Universidade de Vigo (Carmen García Mateo, Eduardo Rodríguez Banga, Leandro Rodríguez Liñares y Xavier Fernández Salgado). Esta colaboración entre la Lingüística y la Ingeniería es imprescindible, ya que la estructura interna del conversor exige conocimientos especializados en cada una de dichas materias.

El método de síntesis escogido para su elaboración fue el de concatenación de unidades pregrabadas, por lo que su arquitectura general está sustentada por dos grandes módulos funcionales: un módulo lingüístico, encargado principalmente de la generación de prosodia, y un módulo acústico que produce la señal de voz sintética. Vemos esto, de manera esquematizada, en la siguiente figura:

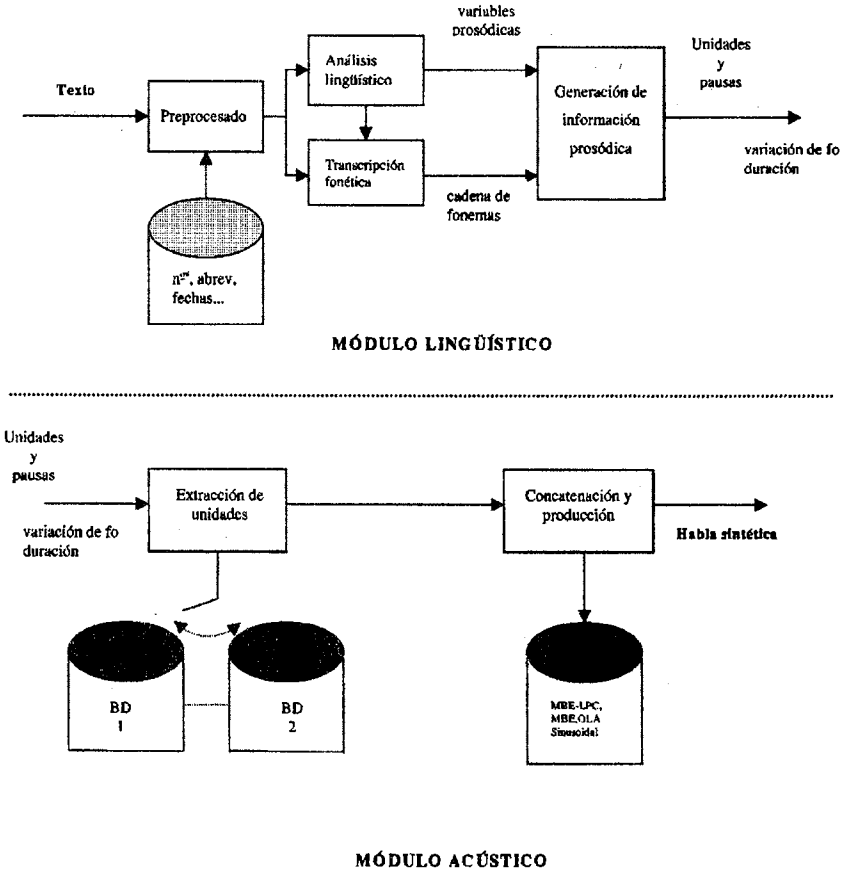

Figura 1. *Diagrama de bloques de un conversor texto-habla basado en la concatenación de unidades*

4.3.1. *Módulo lingüístico*

Este módulo aporta la información lingüística (fonética, morfológica y sintáctica) necesaria para la transcripción fonética y para la generación de la prosodia. Recibe el texto tal y como fue escrito según las normas ortográficas de la lengua gallega y su misión es extraer de él la mayor cantidad posible de información lingüística, para obtener una transcripción fonética completa, tanto segmental como suprasegmental, de modo que esté en disposición de escoger los alófonos adecuados, que formarán la cadena hablada, así como de asignar la prosodia correcta y el pausado adecuado a la frase (curva de entonación, duración e intensidad). Para extraer toda esta información se desarrolla un trabajo siguiendo varias etapas.

En primer lugar, se lleva a cabo la obtención de la frase, ya que el texto al que nos enfrentamos habitualmente es más o menos largo, por lo que, para que su procesado, hay que segmentarlo en unidades más pequeñas. Se considera como unidad óptima a este efecto la frase, término que no designa ninguna unidad sintáctica estrictamente, si bien se acerca al sentido que tiene "oración" en la Gramática Tradicional. Esta unidad aporta normalmente un sentido completo y permite proceder a su estudio aislado sin interferir ni romper la estructura prosódica. Para efectuar esta división del texto en frases, se atiende preferentemente a los signos de puntuación,

solucionando los casos conflictivos que se producen por la aparición de puntos y comas en abreviaturas, siglas y números.

En segundo lugar, se realiza el denominado preprocesado lingüístico. Una vez que se aísla la frase del resto del texto, se lleva a cabo su normalización. Inicialmente realiza una clasificación de palabras (números, horas, palabras que comienzan por mayúscula, abreviaturas...) y, a continuación, hace una extensión de aquellas palabras que lo necesiten:

- desarrolla abreviaturas y siglas, mediante una lista de las más habituales (a modo de diccionario). También posee una pequeña lista para la interpretación de los signos matemáticos.
- interpreta los números, tanto arábigos como romanos y tanto cardinales como ordinales, a través de una serie de reglas que se confeccionan a partir de la lectura de unidades y decenas. Existe otro conjunto de reglas para la interpretación de los números que contempla la posibilidad de que estén representando horas, con la consiguiente diferencia de lectura, sobre todo cuando aparecen minutos y segundos.

Dentro de este proceso de normalización, encontramos otro problema: la correcta lectura de los extranjerismos, ya que la correspondencia carácter-sonido no es la misma que se da en las palabras patrimoniales. Para solucionarlo, se confeccionó otra lista que proporciona una transcripción que trata de representar, empleando la grafía gallega, la pronunciación más próxima posible a la lengua de la que procede la palabra en cuestión.

Hay que tener en cuenta que tanto el listado de extranjerismos como el de abreviaturas y siglas puede ser modificado y personalizado, según las necesidades que se vayan detectando y las aplicaciones para las que se emplee el sintetizador.

Tras la obtención de la frase y su posterior normalización, se lleva a cabo la silabación y acentuación, que son procesos imprescindibles para la correcta asignación del acento prosódico y que se realizan también mediante reglas. La acentuación definitiva consta de dos etapas: una primera donde cada palabra dispone de un acento propio y una posterior donde esta asignación de acentos será modificada, una vez que el conversor posea los datos morfosintácticos que repercuten en dicha acentuación.

A continuación, el sintetizador realiza una primera transcripción teniendo en cuenta los datos que le aportan la grafía, el preprocesador lingüístico y la primer asignación acentual y silabación, y una última y definitiva, aprovechando la información morfosintáctica, donde se tienen en cuenta las modificaciones necesarias para la correcta modulación prosódica y la asignación final de pausas.

Aunque los problemas relacionados con la elección de alófonos se tratarán con más detalle en el apartado dedicado al módulo acústico, hay un par de dificultades que conviene adelantar en este momento: la asignación del timbre correcto a las vocales de

grado medio y la lectura adecuada, como [ks] o como [S] (prepalatal fricativa sorda) de la grafía {x}, ya que no hay marcas ortográficas que señalen cuál es el alófono que le corresponde a cada carácter según el caso. Estos problemas se solucionaron introduciendo reglas heurísticas, si bien, en el caso de asignación de timbres a los verbos, se cuenta con la ayuda del análisis morfológico.

El análisis morfosintáctico constituye una etapa imprescindible dentro del módulo lingüístico, ya que será el que proporcione la información necesaria para la generación de la prosodia. Por una parte, realiza un análisis morfológico, donde se asignan categorías gramaticales a cada una de las palabras, y, por otra parte, lleva a cabo un análisis sintáctico, donde se distinguen los sintagmas que componen la "frase", según las categorías gramaticales establecidas en el análisis morfológico. Estas categorías no se corresponde con las tradicionales, ya que están en función del análisis sintáctico que se va a hacer y este análisis tampoco se corresponde con el tradicional, pues no tiene como objetivo una descripción exhaustiva de la sintaxis del enunciado, sino que su finalidad es obtener la información necesaria para la asignación de pausas y de patrones de entonación y, por lo tanto, hay mucha información sintáctica (y, en consecuencia, morfológica) que resulta irrelevante. Por tanto, el objetivo es establecer la correspondencia entre los "modelos" morfosintácticos y los patrones entonativos.

Dentro de este apartado son resaltables, por un lado, el diccionario que contiene las series cerradas de palabras (artículos, pronombres, indefinidos, demostrativos, posesivos...) y, por otro lado, el analizador verbal. El primero facilita enormemente la categorización morfológica y, por consiguiente, el análisis sintáctico, ya que sirve de gran ayuda para desambiguar las palabras que pueden pertenecer a más de una categoría gramatical. En cuanto al segundo, localiza dentro de cada frase el verbo, considerado el núcleo a partir del cual se puede emprender el análisis morfosintáctico a toda la secuencia.

Por fin, la generación de información prosódica es el resultado del trabajo realizado en todas las etapas anteriores y constituye una tarea clave en el proceso de síntesis de habla, ya que de ella depende fundamentalmente la variedad entonativa y, como consecuencia, la naturalidad del conversor. El adecuado modelado de los parámetros prosódicos (entonación, duración segmental y energía) es imprescindible para que una secuencia suene inteligible, natural y lo menos monótona posible.

Para la generación de la prosodia se utilizan patrones de entonación extraídos a partir de una serie de enunciados naturales, un conjunto suficientemente representativo, que se grabaron con la voz del mismo locutor que donó las unidades acústicas. Estos patrones representan la evolución de la frecuencia fundamental en segmentos (grupos fónicos), de tal modo que el contorno de frecuencia fundamental es finalmente generado por la concatenación de dichos patrones.

No obstante, la prosodia del conversor dista bastante de la natural no sólo por razones técnicas o de ignorancia del funcionamiento de la entonación, sino también

porque en la prosodia natural intervienen, además de los parámetros mencionados, otro tipo de información no lingüística, como la actitud del locutor, la intencionalidad del enunciado, el estado físico y emocional de quien habla... que difícilmente se ven plasmados en un sintetizador de voz que ni siquiera posee un mínimo de información semántica sobre el enunciado que está pronunciando.

Por último, una vez obtenida toda la información lingüística se procede a la definitiva silabación, reasignación acentual, pausado y transcripción fonética final. En este momento el proceso de síntesis llega al módulo acústico.

4.3.2. *Módulo acústico*

El módulo acústico, el segundo gran bloque dentro de la estructura del sintetizador, produce la señal de voz sintética, ya que es él el que almacena las bases de datos (de unidades y de patrones de entonación) y el que contiene los métodos de concatenación de estas unidades y de generación de la prosodia.

La calidad de voz resultante va a depender en gran medida del tipo de transcripción que se realice (más ancha o más estrecha), ya que para que la voz sintética posea la suficiente inteligibilidad y naturalidad debe reproducir, aparte de los aspectos prosódicos de los que ya hemos hablado, ciertas características del habla natural, como, por ejemplo, la coarticulación. Este fenómeno deriva del hecho de que la articulación de un determinado sonido está condicionada por la presencia de otros sonidos (el que le sigue y el que le precede), que hacen que la posición de la lengua y de los labios cambie de uno a otro de forma gradual.

Dado que el sistema de síntesis utilizado es la concatenación de unidades, estas son escogidas de tal manera que se recojan las transiciones entre los sonidos y, por lo tanto, el efecto de la coarticulación, en la medida de lo posible. Puede pensarse que cuanto mayor sea el número de alófonos, mejor es la calidad de voz que se obtiene, sin embargo no es rentable considerar todos y cada uno de los sonidos existentes en gallego, ya que la calidad que se puede conseguir con un número razonable de unidades, no se supera de modo muy palpable con un número mayor de unidades y, sin embargo, el aumento del inventario de alófonos, aumenta considerablemente la complejidad del proceso.

Así pues, la elección de alófonos fue hecha intentando mantener un equilibrio entre estos dos criterios. Se hizo una selección de sonidos pasando por alto muchas realizaciones fonéticas que no parecían muy pertinentes desde el punto de vista perceptual, si bien estaban implícitas en la base de datos acústica.

Todo ello se consiguió, porque la base de datos del sintetizador está formada en su mayor parte por difonos, es decir, por unidades compuestas de dos sonidos consecutivos y que poseen como característica primordial la conservación de las transiciones, producto de la coarticulación. Con estos difonos se elaboró un corpus, donde se recogen todas las posibles combinaciones y que, posteriormente, fue grabado

por un locutor. Para la elaboración de este corpus se hicieron varias pruebas que condujeron a la conclusión de que los sonidos de más calidad son los extraídos de logátomos, o palabras sin sentido. La elaboración de logátomos parte de la clasificación de los sonidos en graves y agudos, característica que depende de la mayor presencia de energía en bajas o altas frecuencias (lo que en fonética articulatoria se traduce en una localización del punto de articulación en una zona más o menos adelantada o atrasada de la cavidad vocal). Teniendo en cuenta esto, se consideran una serie de sonidos como neutros y son estos los que forman el contexto concreto que rodea la unidad a extraer. El acento del logátomo recae en la vocal de la unidad extraída para evitar una relajación excesiva.

Para la extracción de la unidad, teniendo en cuenta el sistema de concatenación utilizado en el sintetizador, se corta cada uno de los sonidos que forman el difono por su parte más estacionaria, es decir, en aquel punto donde su estructura acústica es lo más semejante posible a la que presentaría si no estuviese condicionada por ningún otro sonido. Con esto se consigue que cuando se concatenan dos difonos el punto de unión no presente gran brusquedad.

Algunos sonidos presentan una estructura difícil de dividir (caso de las vibrantes y combinaciones tautosilábicas de consonante más lateral), por lo que se optó por considerar unidades mayores, los trífonos, que suponen mayor calidad, aunque incrementan notablemente el número de unidades (motivo que obliga a considerar estos trífonos sólo en caso de verdadera necesidad).

Así pues, se elaboró un corpus de logátomos en el que se contemplaron todas las combinaciones de unidades posibles que se dan en el habla, tanto en interior de palabra como por fonética sintáctica. Este corpus fue grabado por un locutor, Xosé Luís Freire, que fue seleccionado después de hacer varias pruebas con distintas voces. Las sesiones de grabación fueron realizadas en varias sesiones en la Radio Galega.

En cuando a la base de datos prosódica, se elaboró también a partir de un corpus grabado por el mismo locutor. En primer lugar, se hizo una clasificación de las principales modalidades oracionales, a las que les corresponde un determinado patrón de entonación. El criterio utilizado para tal selección fue el de escoger aquellos que nos parecieron representativos, los que se aparecen con más frecuentemente en el habla. En segundo lugar, tras su grabación, se etiquetaron para su almacenado en la base de datos correspondiente.

Por último, el algoritmo de suma solapada (tipo OLA) utilizado en el sintetizador de gallego permite, por un lado, concatenar unidades obtenidas del habla natural (como acabamos de ver) sin que se den transiciones bruscas y, por otro lado, modificar las características prosódicas de esas unidades adaptándolas a la prosodia deseada.

5. CONCLUSIONES

El sintetizador que acabo de describir brevemente es el primer producto tangible de las Tecnologías del Habla en Galicia. Con todo, la Tecnología del Habla es una línea de investigación abierta en Galicia, que seguramente se va a desarrollar considerablemente y que va a dar muchos más frutos en un futuro no muy lejano.

No quisiera dejar de insistir una vez más en las importantes aplicaciones de todas estas tecnologías y también en la relevancia que van a tener en el proceso de normalización lingüística de nuestra lengua.

Por lo que se refiere a la parte técnica de estos sistemas, el gran reto al que se enfrentan en la actualidad es, sin lugar a dudas, la prosodia, pues en ella reside en gran medida la naturalidad que pueden llegar a alcanzar los sintetizadores de voz. Es en este aspecto en el que están trabajando más en firme la mayoría de los grupos de investigación dedicados a la síntesis de voz. La dimensión del problema es de gran magnitud, pues existe una variedad inmensa de patrones de entonación en el habla natural, y, aunque consiguiésemos guardar en una base da datos una parte muy importante de ellos, íbamos a tener el problema de encontrar las reglas necesarias para que el conversor sepa en qué contextos tiene que usarlos. Tengamos en cuenta que en la entonación entra en juego tanto el conocimiento extralingüístico que poseemos de la realidad, como los múltiples contenidos expresivos que imprimimos a nuestro discurso.

BIBLIOGRAFÍA

Allen, J., M. Hunnicutt, y D. Klatt. 1987. *From text-to-speech: The MITalk system.* Cambridge: Cambridge University Press.

Bailly, G., y C. Benoit, eds. 1992. *Talking Machines: Theories Models and Designs.* Elsevier.

Furui, S., y M. Sondhi. 1992. *Advances in Speech Signal Processing.* Marcel Dekker.

Garrido, J.M. 1996. *Modelling Spanish Intonation for Text-to-Speech Applications.* Tesis doctoral. Departament de Filología Espanyola. Universitat Autònoma de Barcelona.

Keijn, W., y K. Paliwal. 1995. *Speech Coding and Synthesis.* Elsevier.

López Gonzalo, E. 1993. *Estudio de técnicas de procesado lingüístico y acústico para sistemas de conversión texto-voz en español basados en concatenación de unidades.* Tesis doctoral. Dpto. Señales, Sistemas y Radiocomunicaciones. Universidad Politécnica de Madrid.

Marcos Marín, Francisco A. 1994. *Informática y Humanidades.* Madrid: Editorial Gredos.

O'Shaughnessy, D. 1987. *Speech Communication. Humane and Machine*. Adison-Wesley.

Olive, J. 1993. *Acoustics of American English Speech*. Springer-Verlag.

Ramachandran, R., y R. Mammone. 1995. *Modern Methods of Speech Processing*. Kluwer.

Sagisaka Y., N. Campbell, y N. Higuchi, eds. 1997. *Computing Prosody*. Springer Verlag.

Syrdal, A. 1995. *Applied Speech Technology*. CRC Press.

Vidal Beneyto, J. 1991. *Las industrias de la lengua*. Madrid: Fundación Germán Sánchez Ruipérez. Biblioteca del Libro.