Resultados preliminares sobre SLHMM.

J.E. Díaz Verdejo, J.C. Segura Luna, A.J. Rubio Ayuso
P. García Teodoro, Victoria Sánchez Calle
Dpto. Electrónica y Tecnología de Computadores. Facultad de Ciencias.
Universidad de Granada. 18071 - GRANADA (Spain)

Resumen

En este trabajo se propone un nuevo sistema híbrido para el reconocimiento de voz continua que integra HMM y ANN. Dicho sistema se compone de 3 clases de bloques (LVQ, SLHMM y DP), todos ellos redes neuronales, si bien los denominados SLHMM pueden ser interpretados y entrenados de acuerdo con los HMM. Un SLHMM es, básicamente, una expansión en una red con un número fijo de capas de una red neuronal recurrente con una topología conveniente. Se presentan algunos resultados experimentales preliminares que, comparados con los obtenidos a partir de un sistema basado únicamente en HMM, muestran un incremento en el rendimiento del sistema sencillamente debido a la topología utilizada.

1 Introducción

Varios trabajos recientes han puesto de manifiesto las fuertes relaciones existentes entre ambos tipos de Redes Neuronales Artificiales (ANN), concretamente las llamadas Redes Neuronales Recurrentes (RNN) y los Modelos Ocultos de Markov (HMM) [1] [2]. De hecho, estos trabajos han probado que las RNN, con una topología apropiada, son equivalentes a los HMM.

Nuestro objetivo es la utilización de este tipo de redes neuronales como núcleo de un sistema de reconocimiento de voz continua, con la finalidad de mantener la interpretabilidad según HMM. Esto puede permitirnos mejorar el rendimiento de ambos tipos de sistemas (los basados en HMM y los basados en ANN) aplicando técnicas de redes neuronales a los HMM y viceversa. Por otra parte, si se mantiene la interpretabilidad según HMM de uno de los módulos del reconocedor, es posible entrenar sus parámetros con los algoritmos usuales de los HMM, que son más simples que los correspondientes a las ANN.

Para reconocer voz continua utilizando ANN es necesario evaluar las probabilidades de la secuencia a reconocer localmente, debido a la estructura fija de la red. Este problema aparece cuando se utilizan ANN porque no es posible construir un "supermodelo" de la secuencia de voz simplemente concatenando los modelos elementales, tal como se hace en el caso de los reconocedores basados en HMM.

Para solucionar este problema hemos desarrollado un formalismo que puede evaluar el camino óptimo para el conjunto de probabilidades locales. Este formalismo permite diseñar

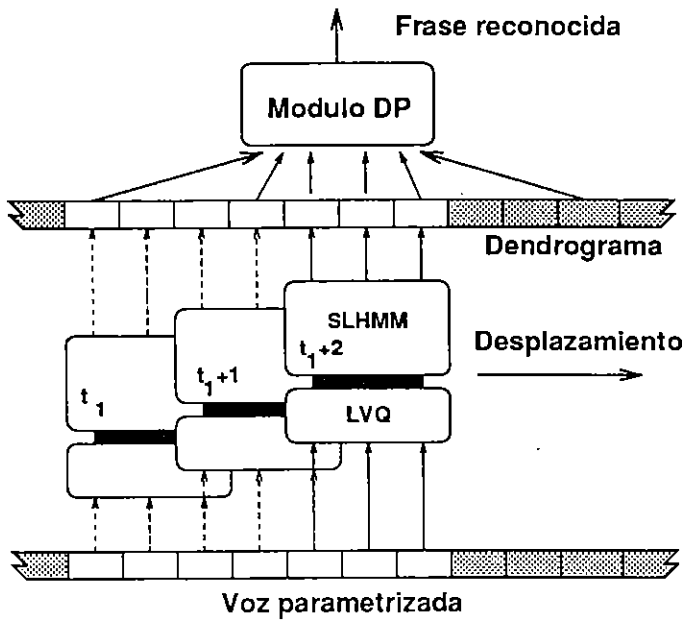

Figura 1: Diagrama de bloques del sistema propuesto.

un sistema modular compuesto por varios tipos de ANN que pueden ser entrenadas mediante los algoritmos convencionales utilizados en ANN, como p.e. retropropagación.

2 Descripción del sistema.

El sistema propuesto está compuesto por 3 tipos de módulos, de los cuales 2 son directamente ANN. El tercer módulo puede ser implementado como una ANN, aunque no es entrenable. Estos módulos son (figura 1): módulos de cuantización LVQ, módulos de estimación de probabilidades locales basado en los SLHMM, y módulo de programación dinámica (DP).

El sistema se compone de un módulo LVQ y un módulo SLHMM asociados entre sí para cada una de las unidades de reconocimiento (en nuestro caso, fonemas) y un módulo de programación dinámica. El procesamiento realizado por el sistema se describe a continuación. La señal de voz es convenientemente parametrizada, obteniéndose una serie de vectores. Esta secuencia de vectores es procesada por cada uno de los módulos SLHMM y LVQ asociados a cada uno de las unidades a reconocer. Para ello se selecciona inicialmente un segmento de longitud determinada de la secuencia de vectores. Este segmento es procesado primero por el módulo LVQ, con lo que se obtiene una versión cuantizada del vector de entrada, que es procesada por el módulo SLHMM. A la salida se obtiene un conjunto de probabilidades parciales del segmento seleccionado a la entrada. Desplazando los dos módulos sobre la secuencia a reconocer es posible obtener todas las probabilidades parciales para todos los instantes de tiempo y para todos los fonemas. El resultado es procesado por el módulo de programación dinámica, que obtendrá la secuencia óptima de fonemas mediante búsqueda sobre el conjunto de probabilidades obtenido anteriormente.

En una primera fase, el módulo LVQ ha sido suprimido debido a que no afecta al núcleo del sistema y, actualmente, nuestro interés reside en el desarrollo de los módulos SLHMM y DP.

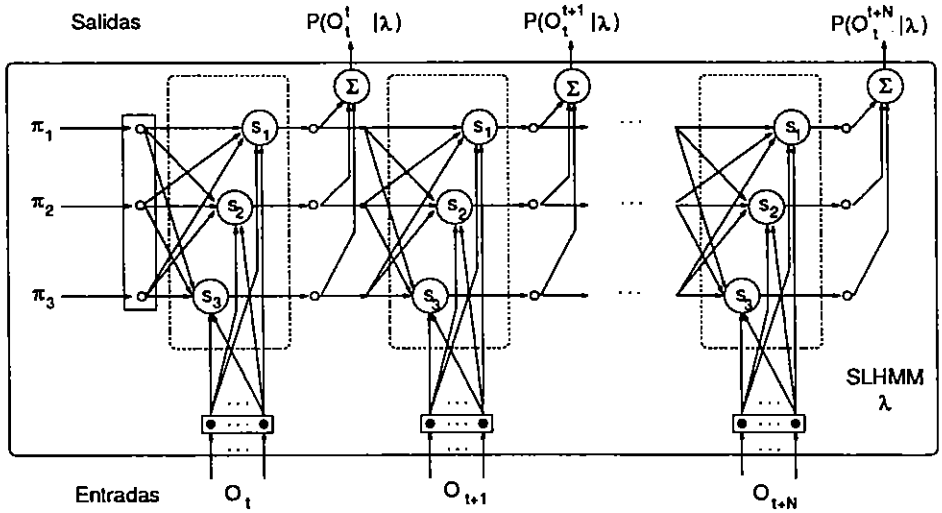

Figura 2: Esquema de un SLHMM.

3 El módulo SLHMM.

El módulo SLHMM es el núcleo del sistema. Básicamente consiste en un desarrollo de una RNN-HMM obtenido expandiendo la red para un número fijo de capas. De esta forma, cada instante temporal en la evaluación de una RNN ha sido sustituida por una capa en la estructura del SLHMM. Obviamente, esta expansión de la red impone restricciones temporales: la longitud máxima de la secuencia a evaluar es el número de capas del SLHMM.

La estructura del SLHMM se muestra en la figura 2. Sus entradas son un número fijo de elementos de la secuencia de símbolos de entrada, mientras que su salida es un vector compuesto por las probabilidades parciales de generación. De esta forma, para un SLHMM con $N+1$ capas, la entrada será la secuencia parcial de símbolos

$$O_{t_1}^{t+N} = \{O_{t_1}, O_{t_1+1}, \cdots, O_{t_1+N}\}$$

mientras que la salida serán las probabilidades de todas las subsecuencias $O_{t_1}^{t_1+N}$ con $1 \leq n \leq N$ y t_1 como instante inicial:

$$\{P(O_{t_1}^{t_1}|\lambda), P(O_{t_1}^{t_1+1}|\lambda), \cdots, P(O_{t_1}^{t_1+N}|\lambda)\}$$

Desplazando la red sobre la secuencia completa (fig. 1) se puede obtener un *dendrograma* (conjunto de todas las posibles segmentaciones) de la señal ya que se evaluarán todas las probabilidades de la forma

$$p(O_{t_1}^{t_2}|\lambda) \quad con \quad 0 \leq t_1 \leq T-N;\ t_1 \leq t_2 \leq t_1+N$$

4 El módulo DP.

El módulo DP obtendrá el camino óptimo sobre el dendrograma. Este camino óptimo viene determinado por el número de unidades que componen la secuencia, L (p.e. el número de

fonemas), la secuencia de unidades óptima, $\{\lambda_1, \lambda_2, \cdots, \lambda_L\}$ y los índices de los instantes temporales finales de dichas unidades $\{t_1, t_2, \cdots, t_{L-1}\}$.

El funcionamieto del módulo DP es diferente dependiendo de si se está entrenando el sistema o reconociendo una secuencia.

Durante el entrenamiento del sistema, las unidades que componen la frase son conocidas. Por lo tanto, el módulo DP realiza un simple alineamiento temporal. Si llamamos

$$\mathcal{L}(t,l) \equiv \log P(O_1^t|\lambda_1, \lambda_2, \cdots, \lambda_l)$$

las ecuaciones de programación dinámica son:

$$\begin{array}{ll} \textit{Comienzo} & \mathcal{L}(1,1) = \log P(O_1^1|\lambda_1) \\ \text{Recursión} & \mathcal{L}(t,l) = \max\limits_{1\leq\Delta t\leq N}\{\mathcal{L}(t-\Delta t, l-1) + \log P(O_{t-\Delta t+1}^t|\lambda_l)\} \\ \textit{Final} & \log P(O_1^T|\lambda_1, \lambda_2, \cdots, \lambda_L) = \mathcal{L}(T,L) \end{array}$$

Este alineamiento puede ser utilizado para el entrenamiento de los módulos SLHMM y LVQ.

Durante la fase de reconocimiento, el módulo DP realiza un proceso muy similar al Viterbi Beam Search debido al hecho de que no se conocen ni el número de unidades que componen la secuencia ni, obviamente, cuáles son dichas unidades. Así, las cantidades a evaluar son de la forma $\mathcal{L}(t,l,\lambda_l)$ dado que dependen del modelo usado. El procedimiento de reconocimiento es el que sigue:

- Dado un valor conocido para $\mathcal{L}(t,l,\lambda_l)$ se evalúan todos los posibles valores $\mathcal{L}(t+\Delta t, l+1, \lambda_{l+1})$ de acuerdo a una gramatica preestablecida

$$\mathcal{L}(t+\Delta t, l+1, \lambda_{l+1}) = \mathcal{L}(t,l,\lambda_l) + \log P(O_t^{t+\Delta t}|\lambda_{l+1}$$

- La inicialización de los valores se hace de la forma

$$\mathcal{L}(1,1,\lambda_i) = \log P(O_1^1|\lambda_i)$$

- Una vez que se ha alcanzado el estado final de la gramatica y el instante de final de la secuencia de entrada, T, se selecciona el valor máximo para $\mathcal{L}(T, L_i, \lambda_{final})$ y se reconstruye el camino óptimo.

5 Resultados experimentales.

El sistema propuesto está siendo evaluado actualmente. Como ya se ha comentado, el módulo LVQ ha sido suprimido en esta fase inicial. Los valores iniciales de los parámetros de todos los SLHMM han sido estimados a partir de un entrenamiento de los HMM de acuerdo con las técnicas habituales para su aplicación a voz continua, con la esperanza de que de esta forma se disminuirá el número de iteraciones necesarias para el entrenamiento del sistema. Los resultados experimentales que se muestran han sido obtenidos a partir de los mismos,

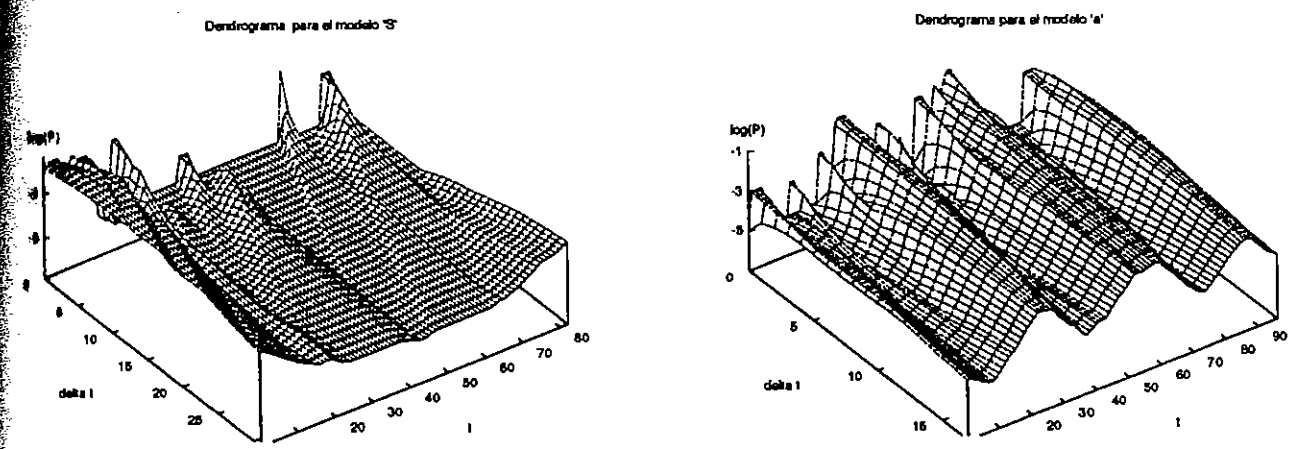

Figura 3: Ejemplos de los dendrogramas obtenidos para varios de los modelos para la frase /dameagua/.

no habiéndose realizado, por tanto, ningún entrenamiento posterior de los modelos mediante las técnicas propias de las redes neuronales.

La evaluación preliminar del sistema se ha realizado sobre un conjunto de 200 frases pronunciadas cada una por 4 locutores (800 frases en total), que han sido divididas en un conjunto de entrenamiento y otro de evaluación. El conjunto de entrenamiento está compuesto por 150 frases, mientras que el de evaluación lo está por las 50 frases restantes. Cada una de las frases ha sido muestreada a 8 kHZ y convenientemente parametrizada de acuerdo a los valores del Cepstrum, ΔCepstrum, Energía y ΔEnergía [3].

En la figura 3 se muestran las probabilidades obtenidas para el fonema /a/ y el silencio en la frase /dameagua/, siendo Δt la duración del intervalo temporal considerado, t el instante inicial del segmento y $log(P)$ el logaritmo de la probabilidad de generación del segmento. En la figura 4 se muestran los valores máximos correspondientes a cada uno de los instantes de tiempo iniciales para todos los fonemas. En dicha gráfica puede observarse el buen comportamiento del sistema, ya que no aparece ningún fonema de los no existentes en la frase con una probabilidad por encima del umbral fijado en la gráfica, y el fonema que localmente proporciona un máximo para la probabilidad se encuentra, por lo general, en la posición adecuada. Unicamente se observan dos pequeños problemas, que son fácilmente eliminados por el módulo de programación dinámica, encargado de proporcionar la secuencia final de fonemas reconocida. Estos dos problemas corresponden a una nasalización de la /d/ inicial (aparece una zona con alta probabilidad para la /m/) y al valor obtenido para la /a/ final de la frase, ya que la el modelo asociado al fonema /u/ proporciona mayor valor para la probabilidad.

Los tasas de reconocimiento obtenidas se muestran en la tabla 1 tanto para el sistema de reconocimiento de voz continua basado en HMM como para el sistema propuesto. En dicha tabla puede observarse que si no se realiza poda (100% de umbral), ambos sistemas proporcionan la misma tasa de error. Esto era de esperar, puesto que los valores de los parámetros de los modelos son los mismos en ambos casos. Sin embargo, para valores similares para el umbral de poda, los resultados sobre la tasa de error son ligeramente favorables para el sistema propuesto. Esta ventaja se hace más evidente si se considera el

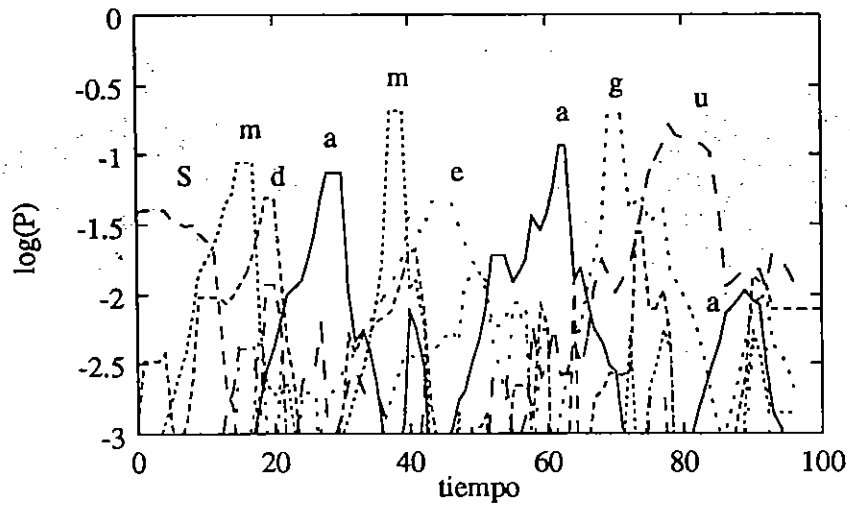

Figura 4: Valores maximos de la probabilidad para los fonemas de la frase /dameagua/.

HMM			SLHMM		
Umbral	Num. nodos	Error	Umbral	Num. nodos	Error
10	8275	34.1	12	3256	32.79
20	16920	9.16	20	13735	8.12
100	—	0.0	100	—	0.0

Tabla 1: Resultados experimentales sobre el conjunto de frases de evaluación.

número de posibles candidatos parciales evaluados (número de nodos), que es bastante más reducido en el caso del sistema propuesto. Esto es consecuencia del propio proceso de poda, ya que en el caso del reconocedor basado en HMM, la decisión de podar o no un camino se hace de forma síncrona, esto es, cada vez que se procesa un vector de entrada, se decide qué candidatos son seleccionados. Por el contrario, en el sistema basado en SLHMM esta poda se realiza de forma asíncrona: la decisión de eliminar o no un posible candidato se realiza atendiendo a las probabilidades de fonemas completos reconocidos.

6 Conclusiones.

Se ha propuesto un sistema de reconocimiento de voz continua que integra modelos ocultos de Markov y redes neuronales cuyo núcleo es el denominado SLHMM. Un SLHMM es una red neuronal que puede ser interpretada según un HMM. El sistema integra también los módulos denominados LVQ que actúan como un cuantizador, con la ventaja de que su estructura permite la utilización de un algoritmo de entrenamiento que optimiza el rendimiento global del sistema. Finalmente, se utiliza un módulo de programación dinámica que obtiene el alineamiento óptimo para la secuencia reconocida.

Un aspecto importante de los SLHMM consiste en la posibilidad que presentan de modelizar la duración de los fonemas. Si durante el entrenamiento de dicho bloque se permite que los pesos de las interconexiones entre las diferentes capas puedan ser diferentes de una capa a otra, el propio algoritmo de entrenamiento obtendrá los valores óptimos de acuerdo

n las posibles duraciones de los fonemas.

eferencias

J. Díaz, "A new neuron model for an Alphanet-Semicontinuous HMM," *Proc. ICASSP'93*, Vol.1, pp. 529–532, 1993.

J. S. Bridle, "Alpha-Nets: A Recurrent 'Neural' Network Architecture with a Hidden Markov Model Interpretation," *Speech Communications*, Vol.9, pp. 83–92, 1990.

J. C. Segura, "Variantes del Modelado Oculto de Markov para Señales de Voz," in *Monografías del Dpto. de Electrónica*, vol. 24. Dept. de Electrónica y Tecnología de Computadores. Universidad de Granada, Diciembre 1991.

AHOZKA: UN SISTEMA DE CONVERSION DE TEXTO A VOZ PARA EL EUSKARA.

I. Hernáez (*), J.C. Olabe (**), P. Etxebarria(***), B. Etxebarria, A.Cuesta(*)

(*) Dpto. de Automática, Electrónica y Telecomunicaciones, U.P.V./E.H.U.

(**) Christian Brothers University, Memphis, Tennessy, U.S.A.

(***) Dpto. Euskal Filologia, U.PV. / E.H.U.

1. INTRODUCCIÓN

En esta comunicación se describe la primera versión de un sistema de conversión de texto a voz desarrollado para la lengua vasca. El esquema básico del sistema se muestra en la figura 1.

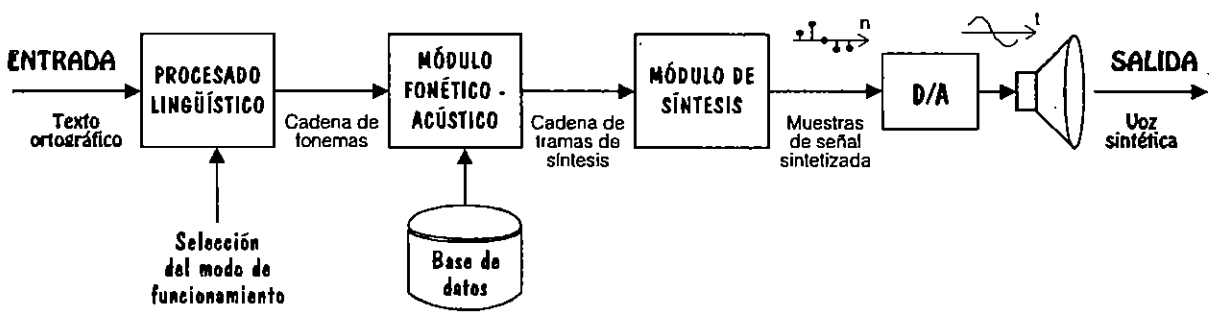

Fig. 1. Diagrama de bloques del sintetizador

El sistema permite la selección de tres modos diferentes de funcionamiento: *modo fonema, modo palabra* y *modo frase* según se quiera procesar cada letra, cada palabra o cada frase de forma independiente. En lo que sigue se explicará el funcionamiento del sistema en modo frase, por ser el más completo e incluir todas las posibilidades del conversor.

El proceso de síntesis comienza con el módulo lingüístico, cuya entrada es texto ortográfico convencional. Este módulo realiza la segmentación del texto de entrada por grupos, dependiendo del modo de funcionamiento elegido. Además de procesar expresiones en formatos comprimidos, tales como fechas o abreviaturas el módulo lingüístico realiza la transcripción fonética del texto y le añade sus características prosódicas. La salida de este módulo es una sucesión de símbolos que representan a los sonidos, junto con la información prosódica necesaria para la síntesis. El funcionamiento detallado de este módulo se explica en la sección 2.

El módulo fonético-acústico podría ser llamado también conversor de sonidos a parámetros, puesto que su función es convertir la cadena de sonidos que tiene a su entrada, en la sucesión adecuada de parámetros requeridos a la entrada del módulo de síntesis, tal y como se describe en la sección 3. Para realizar su función, el módulo fonético-acústico necesita información de cada sonido, información que proporciona la base de datos. Cómo hemos obtenido esta base de datos se describe en el apartado 4.

El módulo de síntesis solicita un número determinado de parámetros cada 10 ms., valor que hemos elegido para la trama de síntesis. El sintetizador entrega al convertidor digital/analógico muestras a razón de 8000 muestras por segundo. El esquema del sintetizador y sus parámetros se describen en la sección 5.

El sistema completo está desarrollado sobre un ordenador personal, programado en lenguaje C. Para la conversión digital/analógica actualmente se utiliza una tarjeta de sonido comercial de bajo coste. El tiempo de ejecución del proceso completo en un 486 a 50 MHz es aproximadamente de 1,5 veces tiempo real, aunque este tiempo puede ser reducido fácilmente, como se explicará en el último apartado.

2. PROCESADO LINGÜÍSTICO

2.1. Introducción

El bloque denominado Procesado Lingüístico, se compone a su vez de cinco subsistemas: la segmentación inicial del texto según el modo de funcionamiento elegido, el preprocesado lingüístico, la segmentación en unidades acentuales, la transcripción fonética y el módulo de análisis prosódico (Fig. 2).

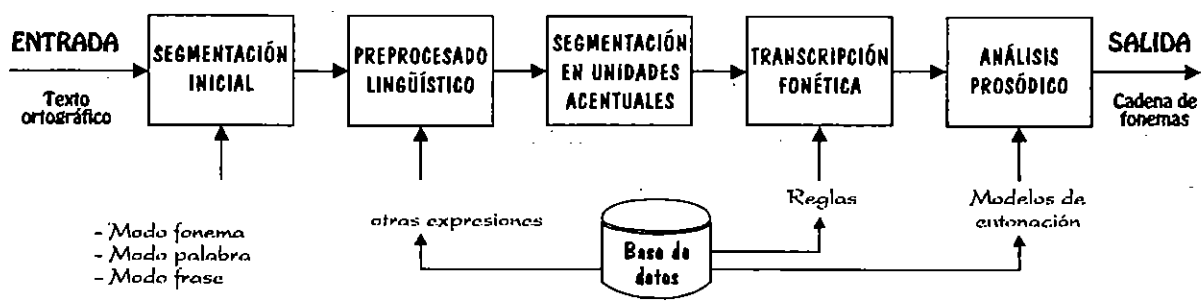

Fig. 2. Preprocesado lingüístico

Un segmentador inicial selecciona la unidad sobre la que se van a realizar las operaciones posteriores. Se explicará únicamente el *modo frase* de funcionamiento. El *modo fonema* y el *modo palabra* simplemente inhiben algunas de los procesos subsiguientes.

El preprocesado lingüístico expande cadenas codificadas en una sucesión de letras. La transcripción fonética transforma estas letras en símbolos asociados biunívocamente con los sonidos. Este proceso de transcripción no puede realizarse a nivel global de frase, como se explica más adelante, por lo que es preciso realizar una segmentación de la frase en unidades menores tales como la *unidad acentual*.

El análisis prosódico asigna a cada sonido valores adecuados de la frecuencia fundamental, la energía y la duración. El algoritmo de generación de las curvas de entonación requiere la silabificación del texto, y la asignación de acentos. Estas dos operaciones deben realizarse sobre la unidad acentual, pero la curva de entonación definitiva, así como el cálculo de las variaciones de la energía y de la duración de los sonidos deben disponer de la frase completa, función que desempeña una memoria intermedia ó *buffer* (Fig.3).

2.2. Preprocesado lingüístico.

El primer paso en el módulo lingüístico consiste en el preprocesado del texto de entrada. En este paso los símbolos del texto de entrada deben ser convertidos a grafías correspondientes a las letras del alfabeto que permitan una pronunciación directa. Se contemplan varios casos:

- Conversión de siglas, abreviaturas y nombres de letras en las cadenas correspondientes extendidas. Esta conversión se realiza por simple sustitución de cadenas tras su detección. La lista de sustituciones realizables se encuentran en un fichero de texto en forma de tabla, accesible al usuario (y por tanto ampliable y modificable en cualquier momento).

- Conversión de fechas y horarios expresados de forma comprimida, y palabras compuestas, con un tratamiento mas elaborado que en el caso anterior.

- Por último lo que hemos llamado preprocesado matemático, con el tratamiento de los números cardinales y ordinales y diversas operaciones matemáticas simples (+ - * / ^), símbolos (<, >, =, <=, >=, etc.) y funciones (sin, cos, tag).

2.3. Transcripción fonética.

La cadena de grafías obtenida tras el preprocesado lingüístico es la entrada al módulo que realiza la transcripción fonética. Existen una extraordinaria variedad de dialectos vascos, y para el desarrollo de nuestro

sistema ha sido preciso elegir un tipo de pronunciación determinada. Aunque existe una normativa precisa para la escritura del euskara (el Euskara Batua, o Lengua Vasca Unificada), para la pronunciación no se dispone de momento de una normativa clara. Sin embargo se han realizado esfuerzos en esa dirección, que han servido de guía para el desarrollo de este trabajo. Debido a la sensibilidad existente en la comunidad vasco parlante con respecto a este tema hemos querido diseñar un sistema abierto a todas las posibles pronunciaciones. Así, las reglas para la transcripción fonética se especifican en un fichero de texto, y mediante una sintaxis muy sencilla, que permite una rápida adaptación del sistema a cada nueva variante.

Una de los aspectos mas interesantes de las reglas de la fonología vasca es su interacción con algunas operaciones morfológicas [Hua-88]. Por ejemplo, un fenómeno generalizado es la palatalización de las consonantes /t/, /l/ y /n/ precedidas de la vocal /i/, es decir (t, l, n) + i -> (c, l, n). Esta palatalización no se da con los sufijos -tar (nativo, habitante de) y -tasun (calidad) (argitasun, y no 'argi*c*asun', y baserritarra, no 'baserri*c*arra'). Tampoco los nombres compuestos sufren la transformación (begi*l*uze -ojos-largos, curioso, y no begi*l*uze). (*-c* es el símbolo utilizado por la I.P.A. para la oclusiva palatal, y l el de la lateral palatal). Esta influencia no es posible tenerla en cuenta si no se realiza un análisis morfológico del texto, lo que aumenta considerablemente la complejidad del proceso de transcripción. Aunque estos fenómenos de interacción no son muy numerosos, sería deseable tenerlos en cuenta. Esto no es posible en la actual versión del sintetizador, puesto que no disponemos de un sistema de análisis morfológico automático.

Además, parece que determinados fenómenos fonológicos entre palabras consecutivas no se dan en cualquier contexto, sino exclusivamente en el interior de los sintagmas [Gaminde, Comunicación Privada]. Así, por ejemplo, 'seme etorri da' daría '*semetorrida*' (ha venido el hijo), pero en 'seme hemetik etorri da' tendríamos vocales geminadas, es decir 'semeemetiketorrida' y no 'sememetiketorrida'. En este caso lo que necesitamos es pues un análisis sintáctico, del cual no disponemos. Puesto que no aplicar la regla (eliminación de vocales geminadas) en un caso en el que habría que aplicarla, parece ser menos grave que lo contrario, aplicar la regla cuando no se debe, la solución está en limitar el ámbito de actuación de las reglas de transcripción fonética al interior de unidades que seamos capaces de aislar, y en las cuales estamos seguros de que estos fenómenos ocurren. Estas unidades son las unidades acentuales, cuyo proceso de formación se explicará en la sección siguiente.

4. Análisis prosódico.

4.1. Introducción

El proceso que sigue a la transcripción fonética, es el análisis prosódico (fig. 3). Los parámetros acústicos que definen la prosodia son la frecuencia fundamental, la energía y la duración. Los valores de la energía y la duración se obtienen en forma de variación sobre el valor intrínseco correspondiente que se asignará en el módulo fonético-acústico. Para la obtención de estos valores se aplican criterios de acentuación, y otros que tienen en cuenta la estructura de la frase.

El valor de la energía queda sólo ligeramente influenciado por el acento en euskara. Así, se aplica un ligero incremento sobre su valor intrínseco si el sonido en cuestión se encuentra en el interior de una sílaba acentuada. También se ha considerado un descenso gradual de la energía en el caso de frases de gran longitud.

En cuanto a la duración de los sonidos, tampoco parece verse afectada por el acento. Actualmente estamos realizando un estudio, en el que se valoran las diferencias de duración de los sonidos en función de sus posiciones relativas en la palabra, en la unidad acentual y en la frase, siguiendo básicamente los criterios seguidos en [Kla- 87] para el inglés, y [San-88] para el español.

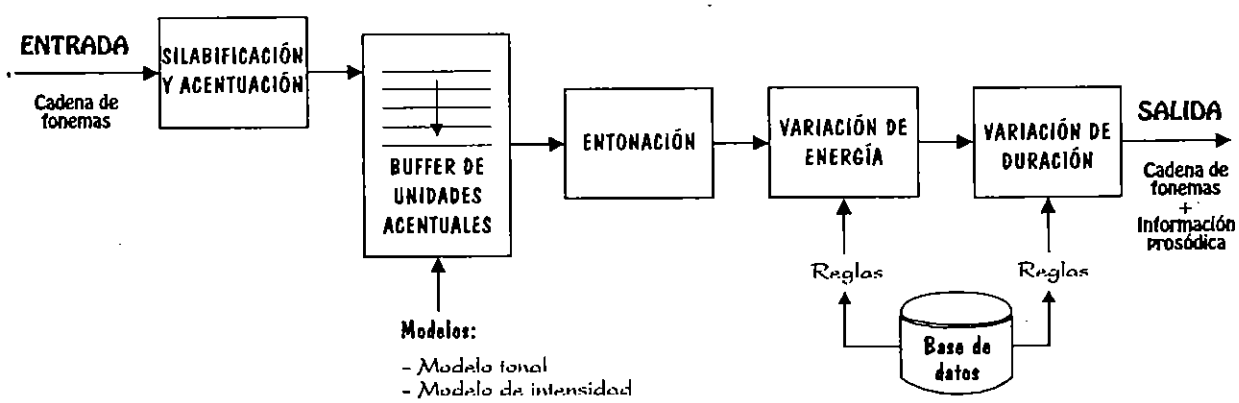

Fig.3. Análisis prosódico

2.4.2. El acento

Como primer paso en el análisis prosódico debemos encontrar las posiciones de los acentos en el interior de cada unidad acentual. El acento es sin duda uno de los temas mas interesantes y también mas controvertidos de la fonología vasca. De acuerdo con la bibliografía consultada, existe una enorme diversidad de mecanismos

acentuales distintos dentro de las diferentes variedades dialectales. Así, es difícil encontrar investigadores que coincidan totalmente en sus conclusiones, debido probablemente al hecho de que no están hablando de los mismos dialectos. De acuerdo con Hualde [Hua-86] en euskara encontramos sistemas acentuales del tipo tonal (*pitch-accent*) en la zona central-occidental, que comprende básicamente las áreas de Vizcaya y Guipuzcoa, junto con sistemas acentuales de intensidad (*stress-accent*) en la zona oriental. Hualde dedica sus trabajos a los acentos vizcaínos, elaborando modelos de generación del acento, cuyas reglas varían según la variedad dialectal. Para el acento vizcaíno disponemos también de los estudios de Gaminde [Gam-93], que además proporciona datos del valor de la frecuencia fundamental. También Txillardegi en [Txi-85] proporciona datos y modelos, aunque su interpretación de los hechos es considerablemente distinta de las anteriores.

Desde un punto de vista de implementación práctica, una dificultad añadida es la inevitable necesidad de realizar un análisis morfosintáctico del texto para aplicar cualquiera de los mecanismos propuestos. Es necesario poder distinguir por ejemplo entre casos y números. También se debe poder extraer la raíz, puesto que ciertas raíces están acentualmente 'marcadas' y no seguirán las reglas de aplicación general después de la sufijación. En todos los casos es necesaria la búsqueda de la unidad acentual, que incluye enclíticos y proclíticos. La formación de esta unidad requiere un análisis sintáctico, puesto que los verbos auxiliares y los artículos determinados, por ejemplo, se comportan como enclíticos. Otros clíticos en cambio son fácilmente detectables, pues son partículas aisladas (ez, baino, bezala...). En nuestro sistema, será posible tener en cuenta únicamente aquéllos clíticos que somos capaces de detectar sin realizar el análisis sintáctico, por lo que la unidad acentual generada será siempre de longitud menor (o igual, en el caso sin error) que la unidad acentual teóricamente correcta.

La solución adoptada en el proceso de asignación de acentos es por tanto una solución subóptima, en la que se han seguido en lo posible, careciendo del tan necesario análisis morfosintáctico, las propuestas de Txillardegi [Txi-80, 87, 92]. Por ello, sobre la *pseudo*-unidad acentual, aplicamos la regla general (+2, -1), es decir acentuación de la primera sílaba de la unidad comenzando por la derecha, y de la segunda comenzando por la izquierda tal y como propone Txillardegui, y *grosso modo* de acuerdo con los demás investigadores.

En lo que se refiere a las palabras marcadas, se ha incluido un pequeño diccionario que detectará los casos sencillos, que no necesiten análisis morfológico. Estas palabras marcadas son transparentes al proceso anterior de asignación de acentos, puesto que no siguen las leyes generales de acentuación.

En cualquier caso, se ha mantenido abierta la posibilidad de asignación de acentos mediante la utilización del acento gráfico que inhibe la localización interna de acentos. .

La posición del acento es el principal determinante de la frecuencia fundamental, y en esto coinciden todos los investigadores (aunque no están de acuerdo en cómo). En cuanto a su influencia en la energía y en la duración, no está, creemos, suficientemente estudiado, aunque los estudios realizados [Her-91] parecen indicar que su influencia es muy pequeña.

2.4.3. La frecuencia fundamental

Hemos desarrollado dos modelos para las curvas de entonación. Con ello se intenta reflejar a grandes rasgos los dos sistemas de entonación mencionados, el sistema tonal y el sistema de intensidad.

Toda sílaba acentuada lleva asociado un valor de la frecuencia fundamental F0 alto. En el modelo tonal de entonación, las sílabas de la misma unidad acentual situadas entre dos sílabas acentuadas llevarán también un valor alto de F0 . El resto de las sílabas llevan asociado un tono bajo. En el modelo de intensidad, las sílabas no acentuadas mantendrán su nivel bajo de F0.

La forma definitiva de la curva de entonación está determinada por la sintaxis. Se ha trabajado únicamente con frases enunciativas e interrogativas simples. Como primera aproximación al problema se han tomado modelos clásicos de entonación [Coo-81][Pie-81]. La curva de entonación está formada por una sucesión de máximos asociados a los tonos altos y mínimos asociados a tonos bajos. Los máximos siguen una línea de declinación cuya pendiente depende de la longitud de la frase de tal forma que los valores de los máximos inicial y final son fijos. La distancia entre máximos y mínimos desciende gradualmente en ambos modelos desde el comienzo hasta el final de la frase.

Para la obtención de los parámetros del modelo se han analizado 40 frases enunciativas e interrogativas. En estas frases se han variado de forma sistemática tres parámetros: la longitud de la frase, desde 8 sílabas hasta 23 (lo que nos ha permitido estimar los valores para los máximos inicial y final); la posición con respecto al comienzo de la frase de la primera sílaba acentuada (para calcular el valor de comienzo) y la posición con respecto al final de la frase de la últimas sílaba acentuada (para calcular el valor final). Los valores intermedios se calculan por interpolación lineal. Se han tenido en cuenta también, como criterio general para la distancia entre valles y picos los datos proporcionados por Gaminde [Gam-93].

3. MÓDULO FONÉTICO-ACÚSTICO

El objetivo de este módulo es la especificación de los parámetros requeridos por el sintetizador para cada trama de síntesis (correspondientes a 10 ms. señal, es decir, 80 muestras a una velocidad de muestreo de 8KHz.), partiendo de los datos disponibles para cada fonema de la cadena de entrada. Dicho de otra forma, debemos convertir la cadena de fonemas en una cadena de tramas de síntesis. Cada trama de síntesis contiene toda la información necesaria para la síntesis. Los datos de partida para cada fonema (o valores intrínsecos de los parámetros) han sido previamente calculados (ver siguiente sección) y se encuentran en una base de datos, con la excepción de la frecuencia fundamental, que procede íntegramente del análisis prosódico.

Algunos fonemas están compuestos por secuencias de sonidos de muy distintas características, (explosivas y africadas) por lo que cada uno de los segmentos que los componen es tratado de forma independiente, disponiendo de información propia en la base de datos.

En la tabla I se han enumerado los parámetros requeridos por el sintetizador. Como algunos de entre ellos suelen tomar un valor fijo, en total, los valores de al menos 17 parámetros deben generarse para cada 10 ms. de señal sintetizada. La tabla II muestra los parámetros que proporciona la base de datos para cada sonido. El resto son parámetros de configuración del sintetizador que pueden mantenerse fijos a lo largo de todo el proceso de síntesis, como se verá en la sección 5.

En el proceso de transformación de fonemas a tramas, el primer paso consiste en la obtención de los valores de los parámetros en la *zona estacionaria* del segmento a sintetizar. Estos valores se obtienen combinando los resultados del análisis prosódico con los datos de la base. Como puede verse en las tablas la base de datos especifica todos los valores en forma directa, salvo las frecuencias de los formantes que se especifican en forma de regla. Como es sabido, las consonantes quedan definidas fundamentalmente por las transiciones de los formantes hacia la vocal siguiente o desde la vocal anterior. Estas transiciones dependen para cada sonido del contexto vocálico en que se encuentre. Por tanto para los sonidos no vocálicos, la base de datos proporciona para cada uno de los tres primeros formantes, el valor que debe ser asignado a un fonema dado, en función del valor del mismo formante en los segmentos de síntesis anterior y posterior, y del contexto vocálico concreto de que se trate.

Una vez definida la zona estacionara del segmento, los valores en las *zonas de transición* se obtendrán por interpolación lineal entre las zonas estacionarias de dos segmentos consecutivos, teniendo en cuenta las duraciones especificadas en la base de datos.

Parámetros de Fuente	
F0	Frecuencia Fundamental
vF0	Varianza de la variación aleatoria de la frecuencia fundamental
BF0	Ancho de banda de la variación aleatoria
glt	Selección Pulso glotal (filtro pas-bajo, pulso de Rosenberg)
n0, n1, n2	Parámetros de configuración del pulso de Rosenberg
BGP	Ancho de banda del filtro paso-bajo (filtro glotal)
FGZ, BGZ	Frecuencia y ancho de banda del antirresonador glotal
BGS	Ancho de banda del filtro paso bajo para excitación quasi-senoidal
AV	Amplitud de la fuente sonora
AVS	Amplitud de la fuente sonora quasi senoidal
AH	Amplitud aspiración
AF	Amplitud fricación
ATT	Contro de variación gradual de ganancia
Parámetros rama serie	
F1C, F2C, F3C, F4C	Frecuencias de resonancia de los formantes
B1C, B2C, B3C, B4C	Anchos de banda de los formantes
FNZ, BNZ	Frecuencia y ancho de banda del antirresonador nasal
FNPC, BNPC	Frecuencia y ancho de banda del resonador nasal
Parámetros rama paralelo	
F1P, F2P, F3P, F4P, F5P, F6P	Frecuencias de resonancia de los formantes
B1P, B2P, B3P B4P, B5P, B6P	Anchos de banda de los formantes
A1, A2, A3, A4, A5, A6	Amplitudes de los resonadores
AN, FNP, BNP	Amplitud, frecuencia y ancho de banda del resonador nasal
Interruptores para conexión de excitaciones	
svc	Conexión excitación sonora a rama serie
svp	Conexión excitación sonora a rama paralelo
shc	Conexión excitación aspiración a rama serie
shp	Conexión excitación aspiración a rama paralelo
snc	Conexión excitación ruidosa a rama serie
snp	Conexión excitación ruidosa a rama paralelo

Tabla I.Parámetros del sintetizador.

4. EL PROCESO DE PARAMETRIZACIÓN

En este apartado se describe la metodología seguida en la obtención de la base de datos utilizada por el módulo fonético-acústico. Para lograr nuestro objetivo, ha sido necesario realizar una parametrización completa de la lengua vasca, y así obtener:

- La duración inherente ó intrínseca de cada sonido (obtenida en un contexto 'neutro').
- La intensidad inherente de cada sonido.
- Las frecuencias y los anchos de banda de los formantes.

Para ello, cada fonema (no vocálico) se ha situado en un contexto intervocálico simétrico (VCV, en donde las dos vocales V alrededor de la consonante C son la misma vocal). La palabra bisilábica así obtenida fue situada en el interior de una frase portadora. Se han considerado los casos oxítono y paroxítono. El método seguido es similar al llevado a cabo para la parametrización del español en [Ola-87]. Para cada palabra se obtuvieron tres realizaciones, por lo que se dispone de un total de 30 realizaciones para cada fonema. Todas las realizaciones fueron obtenidas de un mismo informante, para evitar variaciones en los valores de los formantes debidas a las características del locutor.

Para las medidas de las frecuencias de los formantes se definieron tres puntos de medida para cada uno de los formantes: en las zonas estables de las dos vocales, y sobre la consonante, en un punto de medida adecuado a las características de la consonante. Así, se obtuvieron gráficos como el mostrado en Fig. 4.a. para cada uno de los fonemas, y para cada formante. Mediante regresión lineal, se obtiene el valor que corresponde al formante de la consonante en función de la vocal que le precede o que le sigue. En algunos casos se obtienen dos zonas de comportamiento, como en el mostrado en Fig. 4.b.

Evidentemente, en el habla real la consonante no se encuentra siempre situada entre dos vocales iguales. Para obtener los valores definitivos, los valores obtenidos para las vocales precedente y siguiente se promedian, teniendo en cuenta siempre las duraciones de las transiciones para cada segmento acústico.

De esta misma base de datos fueron extraídos los valores de las duraciones intrínsecas de los fonemas y de los segmentos acústicos que los componen, así como sus energías.

T	Duración de la zona estacionaria
TTA	Duración de la zona de transición anterior
TTP	Duración de la zona de transición posterior
AV,AVS	Amplitudes de la fuente sonora
AF, AH	Amplitudes de la fuente de ruido
A2, A3, A4, A5, A6	Amplitudes de los formantes en la rama paralelo
F1R, F2R, F3R	Reglas para el cálculo de los tres primeros formantes
B1, B2, B3	Anchos de banda de los formantes
FNZ, BNZ	Frecuencia y ancho de banda del cero nasal

Tabla II. Parámetros proporcionados por la base de datos para cada sonido.

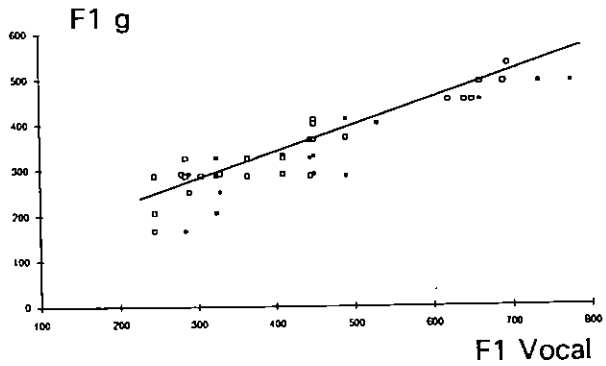

Fig. 4.a . Cálculo de la regla para el primer formante de la consonante 'g'. La ecuación de la recta proporciona el valor de F1 para la consonante, en función del valor de F1 para la vocal.

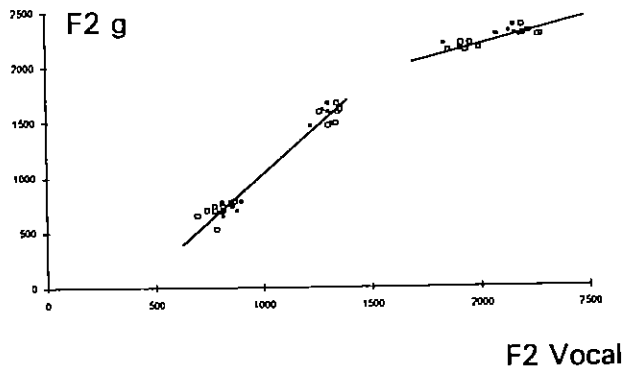

Fig. 4.b. Cálculo de la regla para el segundo formante de la consonante 'g'. Se necesitan dos ecuaciones, segun el rango de valores de F2 para la vocal

5. MÓDULO DE SÍNTESIS.

5.1. Introducción

El sintetizador utilizado está basado en el diseño de Klatt de 1980 [Kla-80]. Se trata de un sintetizador por formantes *serie/paralelo* cuyo esquema básico se ha representado en la figura 5.

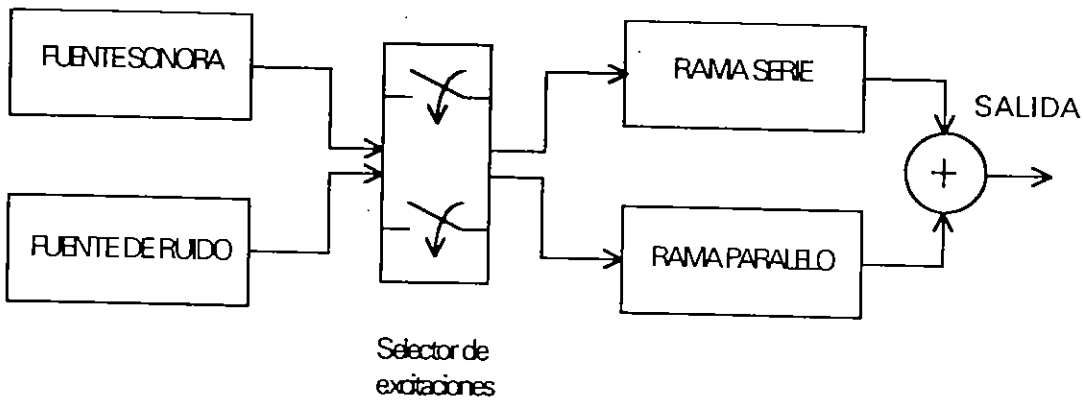

Fig. 5. Esquema general del sintetizador

El sintetizador posee dos fuentes de excitación, la fuente de excitación sonora, utilizada normalmente en la síntesis de los sonidos sonoros, y la fuente de excitación ruidosa, para los sonidos sordos. Como puede verse en la figura 7. ambas excitaciones pueden ser dirigidas hacia las dos ramas, aunque la configuración habitual (svc=snp=1, snc=svp=0, siendo '1' interruptor conectado) dirige los pulsos glotales de la fuente sonora hacia la rama serie y el ruido de fricación hacia la rama paralelo. Una excitación ruidosa dirigida hacia la rama serie permite la síntesis de sonidos aspirados (shc=1, shp=0). Estos puede también generarse si así se desea por la rama paralelo (shc=0, shp=1), aunque no sea lo habitual. La conexión del interruptor svp=1 svc=0, permite utilizar la rama paralelo para la síntesis de sonidos sonoros. Las salidas de las ramas paralelo y serie se suman finalmente para generar la señal sintetizada.

5.2. Excitaciones

Para la generación de sonidos sonoros, impulsos generados con periodo 1/F0 atraviesan una serie de filtros que configuran el pulso glotal. El pulso básico puede obtenerse con un simple filtrado paso-bajo FGP, o bien elegir el pulso de Rosenberg [Rab-78], de calidad ligeramente superior. El antirresonador FGZ, que puede estar desconectado, ayuda a dar forma a la respuesta frecuencial del pulso glotal y conseguir caracterizar distintas voces. El filtro paso FGS se puede utilizar para generar la barra de sonoridad de ciertos sonidos sonoros, pues al ser de ancho de banda estrecho, genera a su salida una señal 'quasi-senoidal'.

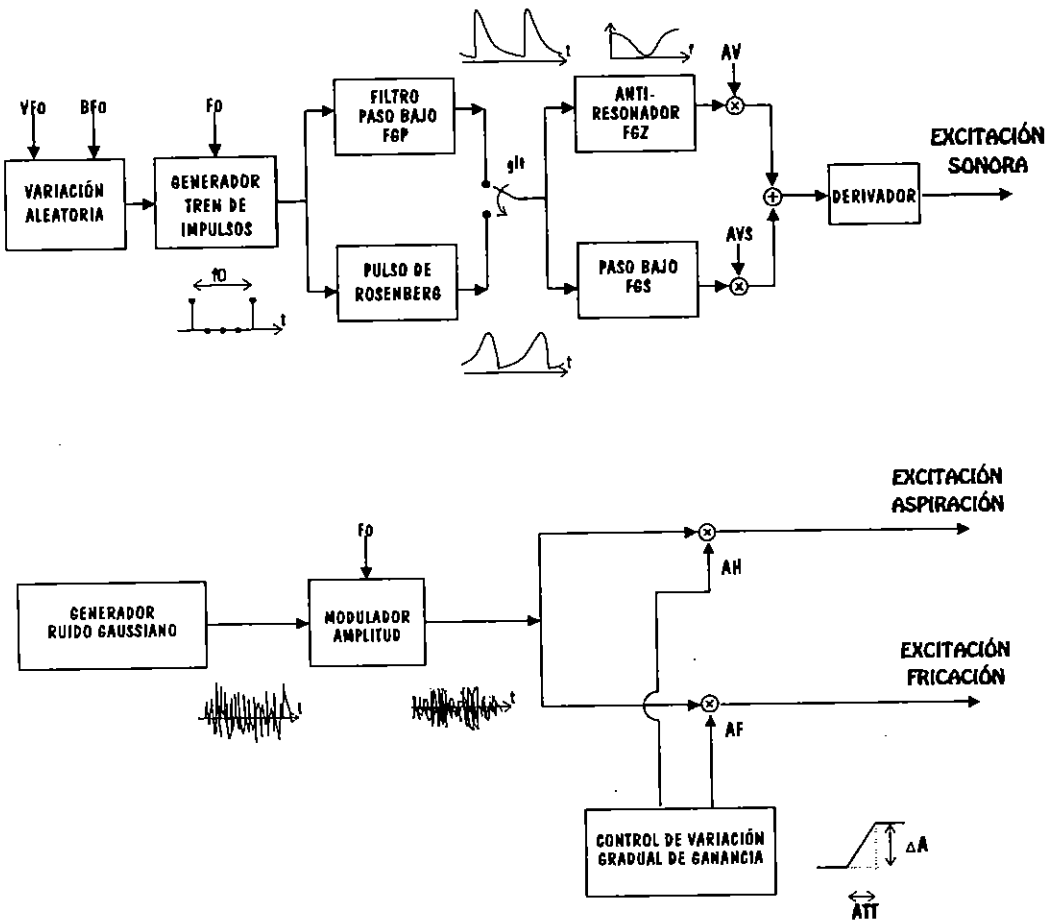

Fig. 6. Generación de la excitaciones del sintetizador.

Para evitar el zumbido resultante cuando se mantiene la frecuencia fundamental constante, hemos introducido una variación aleatoria en dicha frecuencia fundamental. La variación tiene una distribución uniforme siendo posible configurar su varianza (vF0) y ancho de banda (bF0).

La fuente ruidosa utiliza un generador de ruido gaussiano cuya salida puede ser modulada en amplitud por la frecuencia fundamental para obtener las fricativas sonoras. El índice de modulación es constante e igual 0.5. La salida definitiva se obtiene tras aplicar una ganancia, AF si se dirige hacia la rama paralelo para la generación de fricativas, y AH para la generación de ruido de aspiración por la rama serie (shc=1) o por la rama parlelo (shp=1). Para poder controlar la velocidad de cambio de estas amplitudes se ha introducido un control lineal de variación de ganancia, cuya pendiente podemos controlar mediante el parámetro ATT.

5.3. Rama Serie

La rama serie o cascada del sintetizador la forman un conjunto de resonadores y un antirresonador conectados en serie que tratan de simular la respuesta frecuencial del tracto vocal. Los resonadores FxC (F1C, F2C etc.) simulan las resonancias de la cavidad bucal. Es posible utilizar hasta 6 resonadores, aunque para la

frecuencia de muestreo habitual (SR=8000Hz) basta con 4 resonadores. El par resonador/antirresonador FNPC/FNZ, permite la síntesis de los sonidos nasales. Sin nasalidad, el par polo/cero está compensado. Con nasalidad, el cero se desplaza dando lugar a las resonancias nasales.

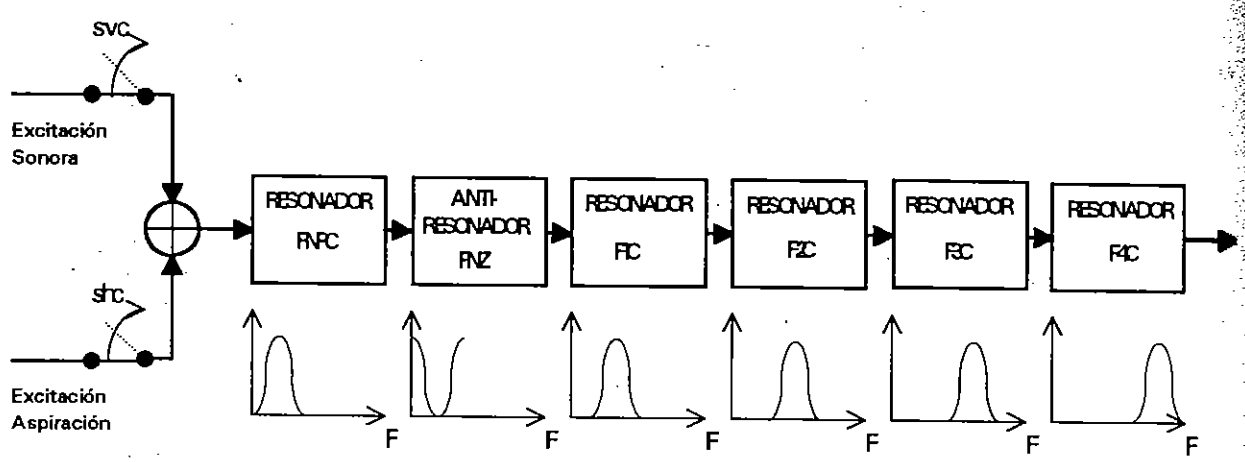

Fig. 7.- Rama Serie o Cascada del sintetizador. La síntesis de sonidos sonoros y aspirador se realiza normalmente por la rama serie (svc=shc=1; svp=shp=0)

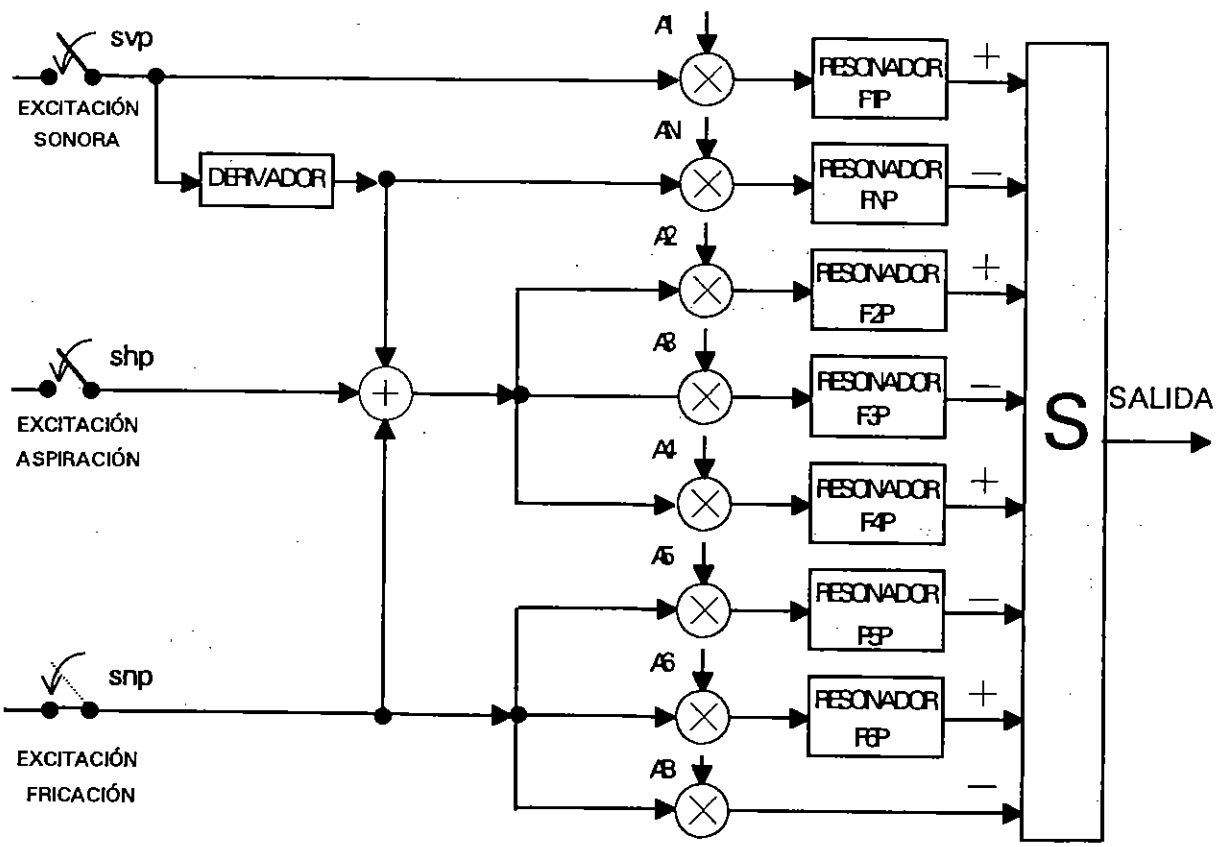

Fig. 8 Rama Paralelo del sintetizador.

5.4. Rama Paralelo

La rama paralelo está formada por una serie de resonadores conectados en paralelo. Para cada uno de ellos podemos controlar la ganancia, y la frecuencia y ancho de banda de la resonancia. En la configuración habitual (svp=0, svc=1, snp=1, shp=0) la fuente ruidosa es la entrada a los resonadores F2P a F6P, que configurarán la respuesta frecuencial para las fricaciones y explosiones. Los filtros F1P (primer formante), y FNPP (resonancia nasal) , se utilizarán únicamente en el caso de generar los sonidos sonoros por la rama paralelo (svp=1, svc=0).

6. CONCLUSIONES.

En esta comunicación se ha presentado la primera versión del conversor de texto a voz para el euskara AHOZKA. La evaluación del sistema no ha sido realizada todavía, pero se ha diseñado ya un corpus con este fin [Etx-92], y está previsto llevarla a cabo en un corto plazo de tiempo.

El sintetizador no trabaja actualmente en tiempo real, pero muy pequeños cambios pueden acelerar el proceso para conseguir tiempo real. Los cálculos se realizan en coma flotante por razones de comodidad. El paso a coma fija no supone un gran esfuerzo, y puede reducir sensiblemente los tiempos de cálculo. Además, el sintetizador se encuentra sobrecargado en cuanto al número de parámetros puesto ha sido la herramienta de diseño del conversor texto-voz, y esta sobrecarga nos ha permitido estudiar distintas posibilidades para la generación de los sonidos. Los programas, particularmente los relativos al módulo de síntesis no están optimizados en cuanto a tiempo de cálculo en pro de una mayor flexibilidad en el estudio de las distintas configuraciones. En cualquier caso, el mercado ofrece microprocesadores cada vez mas rápidos y baratos, por lo que probablemente este problema dejará de serlo en un futuro muy próximo.

Se trata de un sistema muy flexible. Todos los datos lingüísticos se encuentran almacenados en ficheros de texto, lo que permite realizar pruebas de forma rápida. El mismo usuario podría adaptar a su gusto las entonaciones, o las reglas de pronunciación.

Aunque no hemos realizado la evaluación del sistema, creemos que una mejora significativa se puede obtener con la introducción de un módulo complementario de análisis morfosintáctico automático en el módulo lingüístico. Como hemos visto, este análisis mejoraría la 'pronunciación' del sistema por un lado, al mejorar la transcripción fonética, y por otro lado permitiría la obtención de mejores curvas de entonación, puesto que actualmente en múltiples ocasiones los acentos no se introducen en el lugar adecuado, generando una entonación 'extraña'. También está previsto introducir la posibilidad distintas velocidades de elocución, y las variación de las características del locutor, fundamentalmente la introducción de voz femenina.

Bibliografía

1. Azkue, Resurrección M. (1930) *Del acento tónico en algunos de sus dialectos* Euskera IV, pp. 282-318
2. Etxebarria, Pilartxo y Hernáez, Inmaculada (1992) *Test de inteligibilidad para voz sintetizada en euskara* XI Congreso AESLA, Valladolid, Spain
3. Gaminde, Iñaki (1993)*Bizkaiko Azentu-Motei Buruz* Uztaro, 7, 1993
4. Hernáez, Inmaculada y Etxebarria, Pilartxo (1991) *Análisis de la duración de la sílaba en euskara* XXI Simposio de la S.E.L., Granada, Spain
5. Cooper and Sorensen (1981) *Fundamental Frequency in Sentence Production* Springer-Verlag, New-York
6. Hualde, Jose Ignacio (1986) *Tone and Stress in Basque: a preliminary study.* ASJU, XX-3, pp. 867-896
7. Hualde, Jose Ignacio (1987) *A theory of pitch-accent, with particular attention to Basque* ASJU, XXII-3, pp. 915-919
8. Hualde, Jose Ignacio (1988) *A lexical phonology of Basque* Ph.D. Dissertation
9. Hualde, Jose Ignacio (1989) *Acentos vizcaínos* ASJU XXIII-1, pp. 275-325
10. Hualde, Jose Ignacio (1991) *Manuel de Larramendi y el acento vasco* ASJU XXV-3, pp. 737-749
11. Hualde, Jose Ignacio (1992) *Notas sobre el sistema acentual de Zeberio* ASJU XXVI-3, pp. 767-776
12. Klatt, D.H. (1980) *Software for a Cascade-Parallel Formant Synthesizer* J. Acoust. Soc. Am. 67, (3) Mar. 1980, pp. 971-995
13. Klatt, D.H. (1987) *Review of text-to-speech conversion for English* J. Acoust. Soc. Am. 82, (3). September 1987, pp. 737-793
14. Michelena, Luis (1977) *Fonética Histórica Vasca* 2nd. Ed. San Sebastián: Imprenta de la Diputación de Guipúzcoa.
15. J.C. Olabe et al. (1985) *Real time text-to-speech conversion for Spanish.* Proc. Int. Conf. on ASSP, San Diego, CA.
16. Rabiner, L.R. and Schafer, R.W. (1978) *Digital Processing of Speech Signals* Prentice-Hall, Englewood-Cliffs.
17. Txillardegi (J.L.Alvarez Enparanza)(1980) *Euskal Fonologia* Ediciones Vascas, Bilbao
18. Txillardegi (1985) *Euskal Azentuaz* Elkar, Donostia
19. Txillardegi (1987) *Euskara Batua Iruñeko Proposamendua Azentuari Buruz* Udako Euskal Unibertsitatea, Linguistika Saila
20. Txillardegi (1992) *Ahozkera Baturantz* ele aldizkaria,10, pp 65-71