ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA DE TELECOMUNICACIÓ DE BARCELONA (UPC)

Departament de Teoria del Senyal i Comunicacions

SISTEMAS DIFUSOS DINÁMICOS PARA EL TRATAMIENTO DE INFORMACIÓN TEMPORAL IMPRECISA

Autor: Orestes Mas i Casals
Director: Joan Maria Miró Sans

Barcelona, 1997

6.
Aplicación al reconocimiento de voz

6.1 Introducción

6.1.1 El problema del reconocimiento del habla

Desde muy antiguo, el hombre ha utilizado su capacidad innata de hablar como instrumento privilegiado para comunicar sus deseos o intenciones, puesto que es el único método que permite llevar a cabo tal tarea sin un aprendizaje adicional al de la propia lengua materna. Por ello, ya en épocas más recientes, la aparición de todo tipo de máquinas potenció el interés por desarrollar métodos verbales de comunicarse con ellas, en aras de obtener una interacción amigable. Sin embargo no puede hablarse de avances serios en este campo hasta la aparición de las primeras versiones del "espectrógrafo" o "sonógrafo" en las décadas de 1930-40, dispositivo que permite obtener un registro (sonograma) de la energía contenida en las diversas bandas de frecuencia de una palabra o frase en función del tiempo. Al permitir el análisis de la señal vocal de una forma sistemática, el espectrógrafo impulsó las investigaciones en el campo del reconocimiento automático del habla, que culminaron en 1952 con la aparición del primer sistema capaz de discriminar con cierta precisión los diez dígitos ingleses pronunciados de forma aislada por un único locutor [Casacuberta 1987]. A partir de entonces comenzó una explosión de trabajos en esta dirección, con la esperanza de llegar en poco tiempo a la obtención de dispositivos capaces de reconocer de forma precisa todo tipo de frases pronunciadas por una persona cualquiera de forma continua.

Sin embargo, conseguir este hito se ha revelado como una tarea de enorme complejidad, en la que sólo en fechas recientes se han obtenido logros técnico-comerciales aceptables, en parte debido a la aparición de procesadores avanzados a un coste relativamente modesto. Según todos los indicios, las causas del exagerado optimismo inicial se debieron fundamentalmente a una infravaloración de la capacidad de interpretación del cerebro humano en lo que respecta a la comunicación oral. En efecto, estudios cuidadosos del habla han demostrado con posterioridad que ésta constituye en realidad un proceso altamente codificado en el que la voz no es el único agente involucrado, sino que pueden distinguirse múltiples niveles, vías de transmisión y retroalimentaciones. Las características esenciales de este proceso pueden resumirse en los siguientes puntos [Casacuberta 1987], [Terano 1994]:

1. *Bidireccionalidad.* A menudo la comprensión total de un mensaje hablado requiere de una interacción bidireccional entre el emisor (locutor) y el receptor (auditor).

2. *Incompletitud.* La información intercambiada es siempre mayor que la estrictamente contenida en el mensaje oral (gestos, contexto, etc.)

3. *Multiinteractividad.* Existen varios niveles de percepción y/o comprensión que interactúan entre sí, a saber: a) Nivel acústico, en el que se analizan las características físicas de la señal percibida. b) Nivel fonético, en el que se determinan los objetos sonoros elementales (fonemas, sílabas, ruidos simples...). c) Nivel léxico-sintáctico, en el que se extraen los morfemas y se analiza su sucesión para comprobar su adecuación a la gramática del lenguaje. d) Nivel semántico, en el que se llega a la comprensión del significado del mensaje, comprobando la coherencia del mismo y eliminando las posibles interpretaciones absurdas de acuerdo con el contexto en el que discurre el diálogo. Cada uno de estos niveles extrae del mensaje la parte de información que le corresponde para la correcta comprensión del mismo, por lo que es importante que se tengan en cuenta todos ellos en el proceso del reconocimiento automático del habla.

4. *Continuidad.* A pesar de que muchas veces se presuponga lo contrario, no se pueden separar fácilmente de forma automática los distintos elementos constitutivos del mensaje (fonemas, sílabas, palabras...). En el proceso de producción de la voz no se realizan pausas entre ellos y además se influencian unos a otros de tal forma que la pronunciación de un determinado elemento depende de cuáles sean su predecesor y sucesor, fenómeno conocido con el nombre de *coarticulación.*

5. *Variabilidad.* No es posible para una persona pronunciar dos o más veces exactamente igual una misma sílaba, palabra o frase. Aparte

de las variaciones en la amplitud y entonación, uno de los problemas más graves lo suponen las fluctuaciones (de hasta un 50%) en la *longitud* de los vocablos emitidos, las cuales suelen ser además de naturaleza no uniforme. A ello hay que añadir las alteraciones producidas por el estado de ánimo del locutor, condiciones físicas o situaciones especiales (cantar, susurrar, gritar...). El problema se agrava cuando entran en juego varios locutores, puesto que en este caso hay que tener en cuenta que las diferencias de edad, sexo, peso, etc. comportan variaciones en el tamaño del aparato fonador que a su vez redundan en diferencias notables en las frecuencias de resonancia de la cavidad buco-nasal y , por tanto, en el contenido frecuencial de la señal vocal.

6. *Redundancia.* Para transmitir o almacenar de forma perfecta toda la información contenida en una señal de voz se requieren unos 100000 bits/segundo, mientras que hacer lo mismo con el mensaje básico que transporta esta señal requiere tan sólo unos 50 bits/segundo. Esto se puede ver fácilmente si se reflexiona sobre las diferencias existentes entre transmitir una palabra cualquiera en forma de muestras de señal de voz o mediante los caracteres que la componen. El suplemento de información contenido en la señal de voz contiene aquellos elementos que posibilitan identificar al locutor, como la entonación, timbre, etc., y protegen al mensaje de la variabilidad y el ruido ambiente.

7. *Transitoriedad.* La práctica totalidad de los sistemas de reconocimiento actuales se basan en suponer que la estadística de la señal de voz es aproximadamente estacionaria en intervalos cortos de tiempo (unos 20 ms.). Esto es así debido a que el tipo de parámetros que diferencian las transiciones no es aún suficientemente conocido). Este tipo de análisis funciona bien para discriminar fonemas que posean duración, como las vocales y algunas consonantes, pero falla por ejemplo para las consonantes

oclusivas, ya que éstas se producen precisamente en las transiciones. El resultado es que la diferenciación entre consonantes es hoy por hoy uno de los retos más fuertes a superar por estos sistemas.

8. *Incertidumbre e inexactitud.* Generalmente la señal vocal está sometida a toda clase de "artefactos" sonoros que se superponen a ésta, alterándola y dificultando su comprensión. Además, ya se ha mencionado que ésta comprensión sólo se alcanza cuando se han analizado completamente todos los niveles de percepción expuestos en el punto 3, con lo que en general los resultados devueltos por un sistema de reconocimiento que trabaje únicamente a uno de los niveles estarán sujetos forzosamente a incertidumbre.

Así las cosas, en la actualidad se es mucho más consciente que antaño de las enormes dificultades que entraña el proceso del reconocimiento automático del habla, y por ello no suele abordarse este problema desde un único punto de vista. Hoy en día se considera una disciplina propia de la inteligencia artificial, cuyos distintos apartados son adecuadamente enmarcados bajo metodologías de reconocimiento de formas, ayudadas por técnicas de procesado de la señal.

6.1.2 Conjuntos difusos en el reconocimiento de voz. Perspectiva histórica

En los últimos años, el constante incremento de la potencia de cálculo de las CPU especializadas en el proceso de la señal, el gran aumento en la relación capacidad/tamaño de las memorias, así como la notable reducción de los costes de todos estos elementos han provocado que la mayor parte de métodos aplicados al reconocimiento de voz basen su estrategia en la "fuerza bruta" que supone comparar la palabra entrante con un conjunto más o menos extenso de patrones guardados en una memoria. Para conseguir resultados aceptables, estas técnicas deben luchar básicamente con la *variabilidad* de la señal vocal, en la triple vertiente de amplitud, duración y características frecuenciales. Precisamente esta variabilidad es la que ha inspirado a distintos investigadores el aplicar la teoría de conjuntos difusos a la resolución en distintos niveles del problema que nos ocupa.

Históricamente, el primer sistema de reconocimiento del habla que utiliza técnicas difusas es el de Brémont, en Francia [Brémont 1975]. La solución que propone tiene varias fases: en la primera de ellas se procede a captar la señal y a introducirla en un banco de 24 filtros paso banda, cuya salida conjunta constituye una estimación instante a instante del espectro de la señal. Estas 24 salidas se muestrean a 100 Hz y 8 bits/canal, los cuales se comprimen posteriormente a 1 bit/canal mediante un proceso de codificación que asigna un 1 en aquellas bandas de frecuencia en las que se presume la existencia de picos de amplitud, y un 0 en el resto. El resultado de esta fase es una secuencia de palabras binarias de 24 bits que se producen a intervalos de 10 milisegundos. La secuencia completa de palabras binarias producidas al pronunciar una palabra cualquiera delante del micrófono forma de este modo una matriz binaria, que constituye el patrón base para el reconocimiento. Esta matriz presenta unas dimensiones de Nx24, en donde N es igual a la duración del vocablo multiplicado por la frecuencia de muestreo.

En una segunda fase, la matriz binaria anterior se divide en una cuadrícula[18] cuyas celdas constituyen los elementos del universo sobre el que se definen los conjuntos difusos, siendo el grado de pertenencia asignado a cada elemento el número de unos situados dentro de la cuadrícula (grado de ocupación de la misma). Los conjuntos difusos así definidos se guardan en la memoria de un ordenador, y posteriormente son comparados con las palabras a reconocer mediante la utilización de un índice de similaridad adecuado.

La tasa de reconocimiento que Brémont encuentra con este método es del 90%, para un solo locutor y un léxico de 175 palabras aleatorias, con un tiempo medio de cálculo de 1,7 segundos por palabra. Hay que tener presente que en la época de realización de este trabajo la velocidad de proceso de los ordenadores era notablemente inferior a la actual. Para 10 locutores, el sistema sin ningún cambio llega a tasas del 80% para un léxico de 80 palabras. Como referencia, el sistema que utiliza consta de un

---

18 El cálculo exacto de dicha cuadrícula constituye un proceso complejo basado en una segmentación automática del vocablo pronunciado, por lo que no se detalla aquí con profusión. El lector es remitido al texto original para cualquier información adicional.

micrófono, un analizador/codificador, osciloscopio, magnetófono, teletipo/unidad de cinta perforada, alimentaciones diversas y un ordenador T2000 con 8k palabras (de 19 bits) de memoria central. Todo el conjunto ocupa el espacio de una pequeña habitación. Brémont no especifica en ningún momento el coste total del sistema, aunque en la actualidad podría utilizarse perfectamente un ordenador tipo PC con una placa de adquisición para realizar las mismas tareas, con lo que el precio total del sistema se elevaría a unas 100.000 pesetas, cantidad que aún podría reducirse más mediante la construcción de un hardware específico para esta función.

Cinco años después, en 1980, los italianos De Mori y Laface [De Mori 1980] utilizaban un enfoque totalmente distinto, basado en asignar de forma automática etiquetas lingüísticas a los distintos patrones acústicos que aparecen en el habla continua. Estas etiquetas lingüísticas se utilizan posteriormente en estadios más avanzados del proceso de reconocimiento. A grandes rasgos, el sistema tiene la siguiente estructura: La señal vocal de entrada se muestrea en primer lugar a 20kHz y se realiza un análisis espectral de la misma basado en la transformada rápida de Fourier (FFT) y los coeficientes de predicción lineal (LPC). La información obtenida mediante estos subsistemas se utiliza para calcular unas características espectrales "groseras" de la señal (energía total del espectro y en diversas subbandas), las cuales se procesan mediante unas reglas difusas para clasificar los tramos de señal en una serie de categorías. Estas categorías sirven de ayuda a un sistema cuya función es obtener unas características espectrales "finas" de la señal a partir de las FFT y LPC mencionadas anteriormente.

La novedad de este método es, pues, que la clasificación de la señal vocal se realiza mediante la aplicación de reglas y algoritmos difusos, y no mediante simple comparación de patrones. Según los autores, las estrategias de decisión difusa utilizadas por ellos disminuyen un 2% en promedio (del 6% al 4%) la tasa de error de reconocimiento, frente a algoritmos tradicionales basados, por ejemplo, en la distancia euclídea ponderada.

En España también se han construido aplicaciones de reconocimiento de la voz utilizando técnicas difusas, siendo los pioneros en este campo investigadores vinculados al Departamento de Sistemas Informáticos y Computación, de la Universidad Politécnica de Valencia. Su sistema TABARCA-I [Casacuberta 1987] constituye uno de los primeros

intentos de abordar el problema del reconocimiento de forma difusa en todos los niveles de la percepción del habla. El nivel 0 obtiene una descripción simbólico-difusa de la representación paramétrica de la señal vocal. El nivel 1 interpreta difusamente esta descripción en términos de categorías fonéticas groseras, como "vocal frontal" o "fricativa fuerte". Esta última interpretación difusa es procesada en el nivel 2, que a su vez la interpreta difusamente en términos de categorías léxicas. El proceso continúa análogamente en todos los niveles, focalizando los datos de entrada en las interpretaciones más plausibles, pero sin abandonar totalmente las interpretaciones menos evidentes (una característica intrínseca de los sistemas difusos), de forma que los niveles superiores las puedan utilizar si encuentran poca compatibilidad entre las interpretaciones más plausibles y las fuentes de conocimiento introducidas previamente en el sistema.

La característica más sobresaliente de este sistema es quizá que trabaja con una parametrización muy tosca de la señal de voz, con lo cual la complejidad de los cálculos a efectuar se reduce drásticamente. Se utilizan tres parámetros de la señal: su amplitud y la densidad de cruces por cero que presentan la señal y su primera derivada. Estos valores se obtienen cada 15 milisegundos, utilizando una ventana rectangular sin solapamiento. El objetivo de este sistema es el de utilizarse como dispositivo de entrada para una calculadora controlada por voz, por lo que el vocabulario a reconocer consta de los diez dígitos castellanos, mas la palabra "punto", pronunciados de forma continua. Según la referencia mencionada, los resultados que se obtienen a nivel léxico son del 94-99% de aciertos para locutores cuya voz se ha utilizado en el proceso de construcción del aparato, y del 85% para otros locutores. Los autores consideran estas tasas de reconocimiento excelentes dada la crudeza de la representación paramétrica aplicada a la señal vocal.

Un sistema muy parecido al de Brémont es descrito por Terano, Asai i Sugeno en 1989 al hablar de las aplicaciones industriales de la lógica difusa [Terano 1994]. La técnica utilizada en este caso es también la comparación de patrones formados por matrices de bits, aunque en este caso la medida de similaridad utilizada difiere de la de Brémont. Todo el sistema está construido alrededor de un 8086 de Intel, en un equipo de unas dimensiones aproximadamente iguales a las de una hoja DIN-A4. Para un vocabulario de 120 palabras pronunciado por varios locutores, las tasas de

reconocimiento que se consiguen están alrededor del 93%. El sistema admite entrenamiento, y utilizándolo en un único locutor se llega a promedios del 98-99% en la tasa de reconocimiento.

Más modernamente, un sistema de reconocimiento difuso de la voz que utiliza una técnica interesante es descrito por Constantin von Altrock en su libro *Fuzzy logic and neurofuzzy applications explained* [von Altrock 1995]. Básicamente, el método consiste en muestrear y almacenar en una memoria la señal, después de lo cual se considera a ésta como si estuviese representada en la pantalla de un osciloscopio, es decir, en una gráfica amplitud-tiempo. La citada "pantalla" se compartimenta entonces en casillas, cada una de las cuales se considera un elemento perteneciente al universo del discurso de la palabra emitida. Posteriormente se asigna a cada elemento-casilla un grado de pertenencia que es proporcional al grado de "actividad" que la señal vocal presenta en esa zona. Con toda esta información se procede a realizar una inferencia difusa en base a una serie de reglas que relacionan actividades en casillas con palabras emitidas.

Es sencillo percatarse del hecho de que esta técnica permite acomodar variaciones temporales en la señal de voz, además de las variaciones en la amplitud, por lo que efectúa un tratamiento difuso del tiempo, existiendo en este sentido cierta relación con el método propuesto en esta tesis. Sin embargo, las similitudes terminan aquí, pues es obvio que en la aplicación descrita no se pretende sistematizar el tratamiento difuso del tiempo, sino que se busca una solución particular a un problema particular. Esto no obstante, la propuesta tiene un interés notable, y en cierta medida ha inspirado la aplicación de reconocimiento de voz que presentamos en este capítulo.

6.1.3 Ventajas de un enfoque temporal difuso

En la sección anterior ha quedado establecido que la teoría de conjuntos difusos constituye un enfoque perfectamente válido para abordar el problema del reconocimiento de la palabra en cualquier nivel de abstracción. Según destacan los autores en sus respectivos trabajos, los resultados obtenidos por los métodos difusos en comparación con otros más clásicos muestran que, en igualdad de condiciones, las tasas de reconocimiento que presentan los primeros son superiores a las obtenidas por los segundos. Además, de forma similar a lo que ocurre con las redes neuronales, la solución

difusa permite un alto grado de paralelismo, por lo que según la implementación que se lleve a cabo pueden conseguirse velocidades de proceso muy superiores a las alcanzadas por otros métodos. También es importante constatar que pueden obtenerse resultados satisfactorios utilizando únicamente parametrizaciones muy toscas de la señal de voz, cosa que permite una reducción de costes y abre de paso una puerta a la aplicación del reconocimiento del habla en campos en los que antaño le estaba vedada, como la juguetería y en general todos aquellos productos en que el coste sea un factor clave para su éxito comercial.

Sin embargo, todos los métodos desarrollados en el pasado presentan la característica común de enfocar el problema desde un punto de vista digital, pues requieren guardar en memoria el patrón formado por la voz parametrizada. Esto dota al sistema de enorme flexibilidad y posibilita la aplicación de algoritmos impensables en una solución analógica, pero dificulta su utilización comercial en productos económicos puesto que el *hardware* digital mínimo tiene ya un coste excesivo para estos casos. Esto es así porque, si bien es cierto que hoy en día existen en el mercado microcontroladores de propósito general muy completos y económicos, las particularidades del problema del reconocimiento del habla exigen de ellos que dispongan de capacidades aritméticas avanzadas, como operaciones de multiplicación y división, de cantidades generosas de memoria para almacenar los patrones a reconocer, así como de una velocidad de proceso aceptable para poder dar resultados en tiempo real.

Una de las razones más poderosas para adoptar la solución digital ha sido el hecho de que ésta permite compensar las fluctuaciones temporales de la señal de voz. En efecto, desde que en la década de los 50 se empezó a abordar de forma sistemática el problema del reconocimiento automático del habla, uno de los obstáculos que primero se presentaron fue el de compensar la *duración* e *instante de pronunciación* variables de un mismo vocablo pronunciado en distintos momentos o por distintas personas. Para minimizar el problema se introdujo en 1960 la llamada *técnica de normalización temporal no lineal*, que ajusta la duración de los vocablos a un valor constante para poder compararlos entre sí. El inconveniente de esta técnica es que requiere la ejecución de algoritmos más o menos complejos (programación dinámica, etc.) que impiden su implementación mediante técnicas analógicas, más económicas y rápidas que las digitales.

Es por ello que el método de inferencia temporal difusa desarrollado en esta tesis puede constituir un enfoque apropiado en este ámbito concreto, pues permite acomodar de forma natural esas fluctuaciones temporales de las señales, sin necesidad de recurrir a los algoritmos mencionados. Además, las implementaciones a que se llega pueden ser totalmente analógicas, lo que constituye una aportación novedosa en este campo. La clave de todo el proceso reside evidentemente en la utilización de conceptos temporales difusos en las reglas de inferencia. Seguidamente se expone una aplicación de reconocimiento de la palabra que presenta todas estas características. No se pretende con este sistema conseguir ningún logro espectacular en cuanto capacidad de reconocimiento, sino únicamente ilustrar las posibilidades del método de inferencia temporal difusa propuesto en cuanto a facilidad de diseño y competitividad de la solución obtenida. Por ello, se procederá a resolver un problema sencillo: discriminar entre los vocablos SÍ y NO, pronunciados por distintos locutores. El dispositivo deberá encender un diodo LED verde cuando se pronuncie un SI y un LED rojo cuando se pronuncie NO. El comportamiento del sistema ante la pronunciación de otros vocablos no se especifica, aunque es deseable que en estas situaciones no se active ninguno de los dos LED. Aún aceptando el carácter casi exclusivamente académico del problema, hay que destacar su doble atractivo de permitir llegar a resultados concretos en muy poco tiempo y ser fácilmente generalizable a vocabularios algo más extensos y comercialmente interesantes.

El objetivo final perseguido es llegar a la construcción física de un prototipo capaz de realizar automáticamente las tareas prescritas, sin más intervención del operador que la conexión del aparato. La descripción del sistema se efectuará por bloques, procediendo a la simulación previa mediante ordenador del funcionamiento de cada uno, para posteriormente pasar a su realización circuital, ajuste y verificación. A continuación se describe la estructura general del sistema y después se tratan por separado las distintas partes que lo componen. Cabe destacar por último que parte de los resultados aquí presentados se exponen también en otro trabajo previo del autor [Mas 1996].

6.2 *Estructura general del sistema*

Como todo problema de reconocimiento de formas, el que nos ocupa puede abordarse siguiendo el clásico esquema por niveles utilizado en estos casos, y que se muestra en la Figura 6.1 [Lin 1996]:

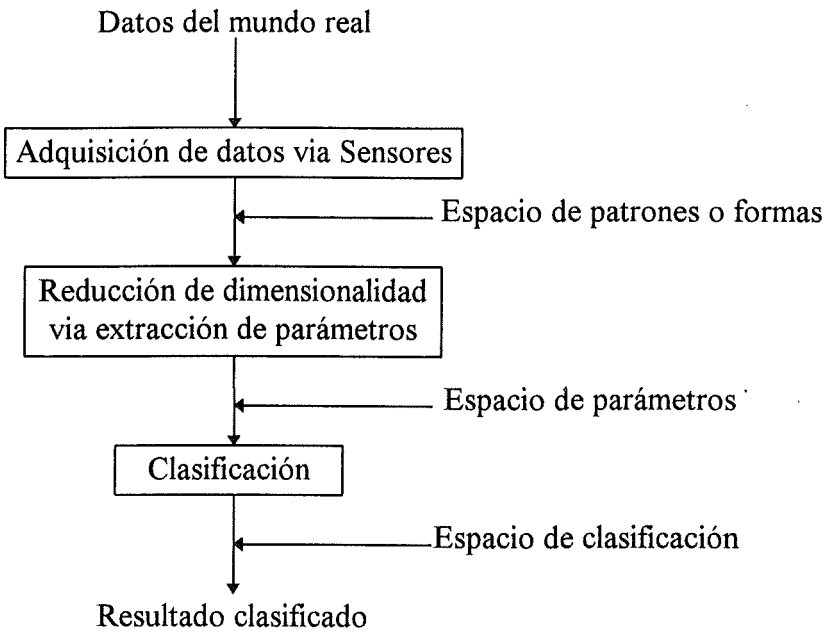

Figura 6.1: Esquema general del reconocimiento de formas

La aplicación de este diagrama de bloques al discriminador que se pretende diseñar puede hacerse de la siguiente forma: En una primera etapa se capta la señal de voz y se encuentra una adecuada parametrización de la misma. Tal parametrización debe ser lo suficientemente rica para permitir la discriminación deseada, pero a su vez lo suficientemente tosca para no complicar innecesariamente las etapas posteriores del sistema. Posteriormente los parámetros obtenidos se utilizan para **clasificar** el fonema que se está produciendo en cada instante en cinco categorías básicas: Silencios (/-/), Nasal (/n/), Vocal débil (/o/), Vocal fuerte (/i/) y Silbante (/s/). Con esta información, puede construirse un sistema de inferencia temporal difusa que discrimine entre los vocablos "SI" y "NO" mediante las dos reglas siguientes:

- **Si** (voz ANTES era /s/) **y** (voz AHORA es /i/) **entonces** (salida es SÍ)
- **Si** (voz ANTES era /n/) **y** (voz AHORA es /o/) **entonces** (salida es NO)

La salida del sistema pueden ser dos tensiones, v_1 y v_2, asociadas a los vocablos SI y NO, respectivamente, de manera que mediante unos comparadores sea posible encender los LED en la forma deseada. Es importante señalar que no se emplea ningún tipo de memoria, por lo que el reconocimiento se realiza en tiempo real. En las secciones siguientes se estudian más detalladamente cada uno de los procesos mencionados.

El diagrama de bloques del sistema discriminador se muestra en la Figura 6.2, y lo constituyen los subsistemas siguientes siguientes:

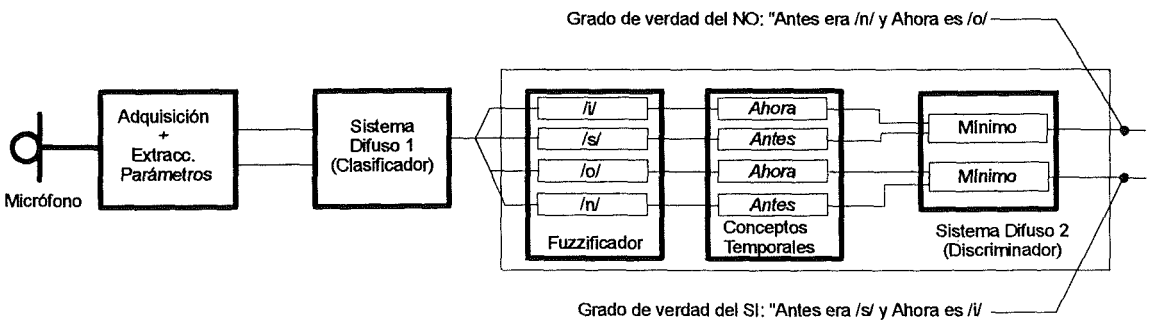

Figura 6.2: Diagrama de bloques del discriminador de vocablos

6.3 Adquisición y preprocesado de la señal vocal

En todo sistema de reconocimiento automático del habla, la primera acción a realizar es la captación y preprocesado de la señal vocal. Esta etapa tiene una importancia clave para las posteriores, pues algunos métodos de parametrización son muy sensibles a los errores introducidos en esta fase, como pueden ser la adición de ruido y *offset*. Para ayudar al diseño es conveniente conocer bien las características temporales y frecuenciales que presentan las palabras a reconocer, por lo que a continuación se procede a capturar y procesar la señal de voz para estudiarla en estos dos dominios. En primer lugar, la Figura 6.3 muestra las gráficas amplitud-tiempo de los vocablos "SI" y el "NO". Las señales se han obtenido muestreando la voz a 22kHz y 16 bits, mediante una placa digitalizadora conectada a un ordenador PC, la cual devuelve un fichero con muestras en el rango $[-2^{15},2^{15}-1]$, las cuales se han normalizado al intervalo [-10,10] para representarlas. En estas gráficas se distinguen con bastante claridad los tramos de señal correspondientes a cada uno de los cuatro fonemas que forman estas dos palabras, a saber: /s/, /i/ y /n/, /o/, así como los silencios que las preceden y siguen, aunque no es posible decidir con claridad los límites de cada fonema puesto que las

transiciones entre ellos son *graduales*. Esta característica es la que permite modelar sus duraciones de forma difusa. A pesar de esta imprecisión, puede observarse que todos los fonemas tienen aproximadamente la misma duración, que es posible describir lingüísticamente mediante el número difuso "unos 150 milisegundos".

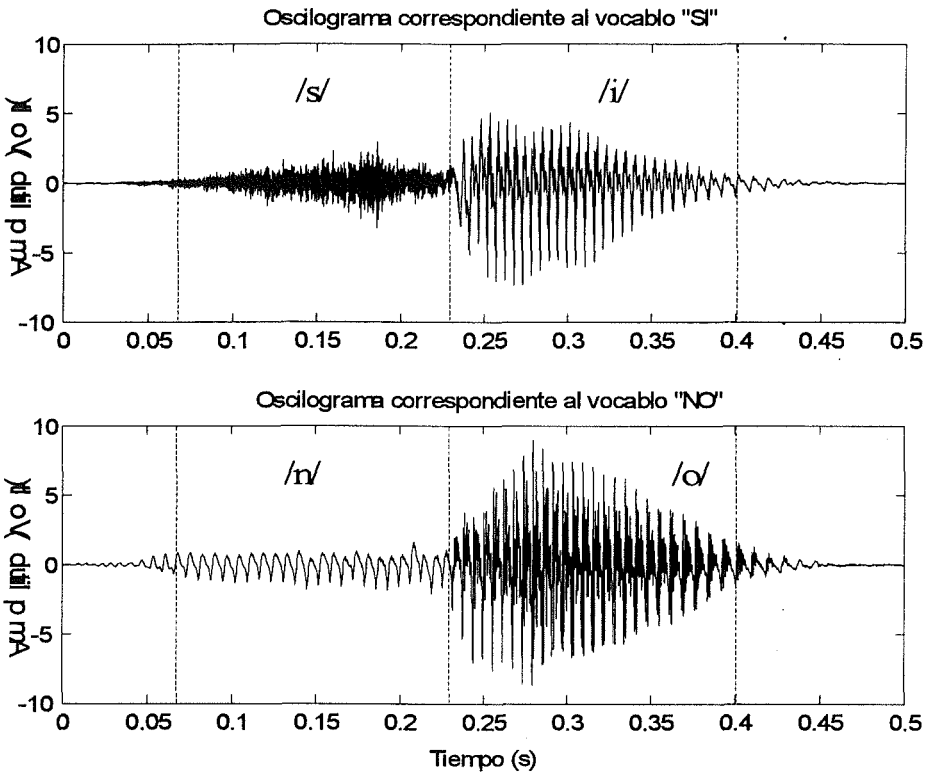

Figura 6.3: Oscilogramas de los vocablos SI y NO

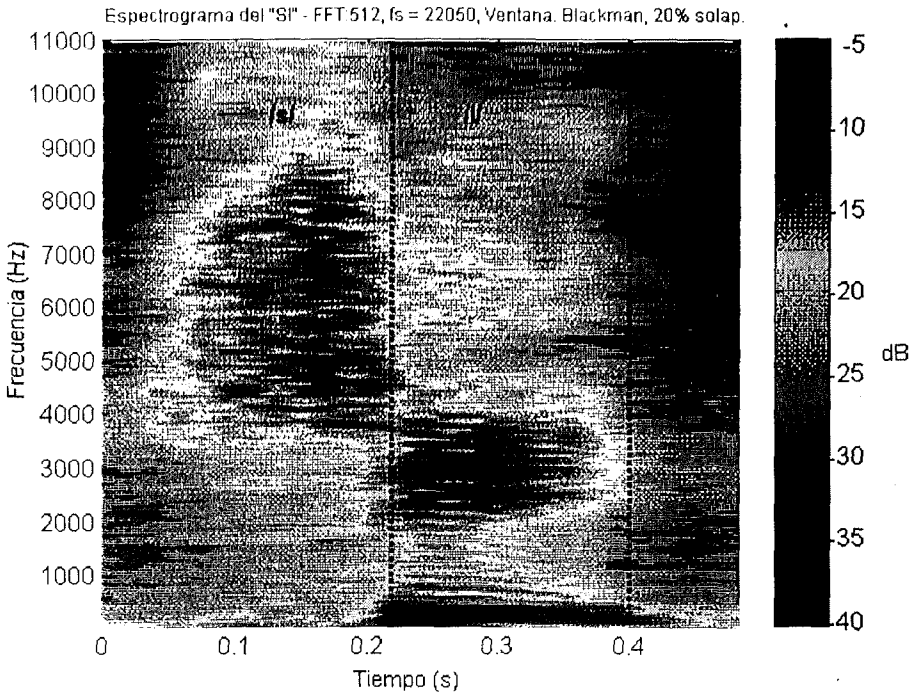

Figura 6.4: Espectrograma correspondiente al vocablo SI

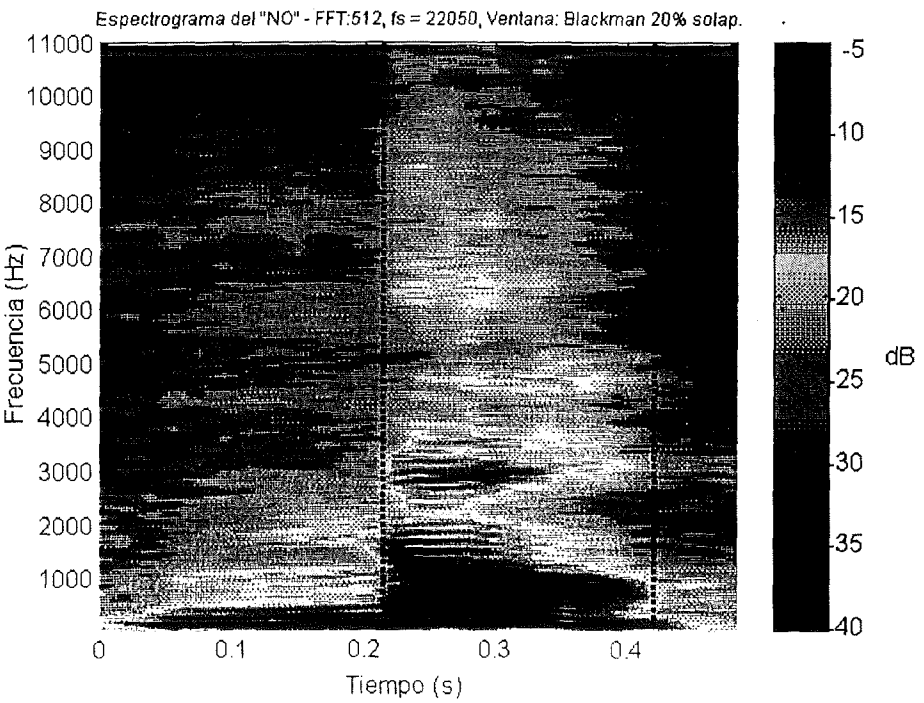

Figura 6.5: Espectrograma correspondiente al vocablo NO

Las gráficas de la Figura 6.4 y la Figura 6.5 corresponden respectivamente a los espectrogramas de las señales temporales de la Figura 6.3. El cálculo de estos

espectrogramas se ha llevado a cabo mediante una FFT de 512 puntos, que a la frecuencia de muestreo empleada en la captura supone dividir la señal en tramos de 23.2 milisegundos; a cada tramo se le ha aplicado un enventanado de Blackman con un solapamiento del 20% entre tramos consecutivos.

Del análisis de estas gráficas pueden desprenderse las siguientes conclusiones:

- El ancho de banda máximo de la señal puede fijarse en 8kHz.
- El fonema /s/, dado su carácter ruidoso, presenta actividad a partir de 4kHz, mientras que las vocales /i/ y /o/ la presentan *hasta* 4kHz. La nasal /n/ tiene un contenido frecuencial muy concentrado en la banda más baja del espectro, llegando a extenderse sólo hasta 500 Hz.
- A simple vista es sencillo distinguir la /s/ de la /n/, y éstas de la /i/ y /o/. Lo que ya es más complejo es distinguir la /i/ y la /o/ *entre sí*.

Esta información será de utilidad en la síntesis de etapas posteriores. De momento, dadas las características temporales y frecuenciales de las señales a reconocer, la etapa de adquisición se ha diseñado de la siguiente forma: se capta en primer lugar la señal de voz mediante un micrófono *electret*, la cual se introduce a un preamplificador que eleva su nivel y le aplica un filtrado paso banda con frecuencias de corte de 100 Hz y 6kHz, con la triple intención de eliminar la componente continua, las posibles interferencias de 50Hz, y de limitar el ancho de banda de la señal resultante. Es importante evitar la saturación del preamplificador o de cualquiera de las etapas que seguirán, ya que ésta provocaría la aparición de componentes de alta frecuencia que falsearían los valores posteriores de los resultados. Por ello, en aplicaciones donde el nivel de señal presente a la entrada no esté predeterminado resulta de interés la incorporación de un circuito de control automático de ganancia (CAG) en algún punto del camino de la señal. La salida de esta etapa la constituye, pues, la señal de voz limitada en banda y con el nivel adecuado.

6.4 Parametrización

Después de la etapa de adquisición y preprocesado, debe aplicarse a la señal de voz un proceso interpretativo para extraer de ella aquellos parámetros que mejor la caractericen. Durante este proceso, el objetivo primordial es reducir notablemente la cantidad de información a manejar mediante la eliminación de las redundancias presentes en la señal [Casacuberta 1987]. Debemos, pues, plantearnos elegir un método de parametrización adecuado a nuestras necesidades.

Sea cual sea la parametrización escogida, deberá ser a la vez *sencilla de obtener y manejar* (por motivos económicos) y *significativa* (suficiente para permitir alcanzar los objetivos de reconocimiento propuestos). Para ello se ha escogido como parámetro la *densidad media de cruces por cero* de la señal de voz en un cierto período. A pesar de ser una representación bastante "tosca" de la voz, dicho parámetro nos proporciona una estimación aproximada del contenido frecuencial de la señal [Ito 1971], [Rausell 1984], suficiente para nuestros propósitos. Para aumentar la información disponible se ha decidido calcular este parámetro dos veces, una para las componentes paso-bajo de la señal de voz y otra para las componentes paso-alto. Así, a la salida del preamplificador se aplican a la señal dos filtros en paralelo, uno paso alto y otro paso bajo, para lo cual se han utilizado unos filtros Butterworth de cuarto orden, siendo la frecuencia de corte de ambos de 2kHz. El valor de esta frecuencia se ha elegido a partir de la observación del espectrograma de la Figura 6.4, en el que puede observarse que una parte importante del contenido frecuencial del fonema /i/ empieza precisamente alrededor de esa frecuencia. Esto no es así para el fonema /ó/, lo que permite suponer que separar las bandas en ese punto debe permitir distinguir los dos fonemas indicados. También es de destacar que el orden de los filtros no se ha escogido excesivamente elevado para que la separación de las bandas frecuenciales se haga de una forma difusa.

Después de los citados filtros, cada una de estas dos componentes se lleva a un circuito detector de cruces por cero, y la salida de éstos se lleva a sendos contadores digitales. La lectura periódica (y posterior puesta a cero) del valor contenido en los citados contadores nos proporciona la parametrización requerida. El período durante el cual se acumulan cruces por cero se ha tomado de *10 milisegundos*, por ser un intervalo de tiempo durante el cuál las características de la señal vocal pueden considerarse

aproximadamente estacionarias [Scarr 1968]. Es de destacar que la detección de paso por cero se realiza con algo de **histéresis** (concretamente el 1%) para conseguir una mayor inmunidad frente al ruido, algo a lo que es muy sensible la parametrización escogida. La Figura 6.6 y la Figura 6.7 muestran los pasos por cero en la banda paso-bajo y paso-alto de los vocablos SI y NO, respectivamente.

A simple vista parece posible distinguir entre los distintos fonemas que componen estos vocablos utilizando únicamente los pasos por cero de la señal filtrada, con lo que parece posible construir un sistema difuso que efectúe la discriminación requerida a partir de estas señales. Es de destacar que para discriminar correctamente es esencial tener en cuenta el *orden* en que se producen los fonemas, pues de otro modo produciría el mismo efecto pronunciar un SI que un IS. En este punto es donde podrían emplearse conceptos temporales difusos para decidir entre una de las dos opciones.

Dado que, como se verá en la sección siguiente, se requieren como mínimo 5 reglas con 2 antecedentes cada una para distinguir correctamente los 5 fonemas involucrados, si asociamos un concepto temporal distinto a cada uno de los antecedentes resultan un total de 10 conceptos temporales a diseñar. Aunque esto es perfectamente abordable, se ha preferido efectuar previamente una nueva reducción de dimensionalidad

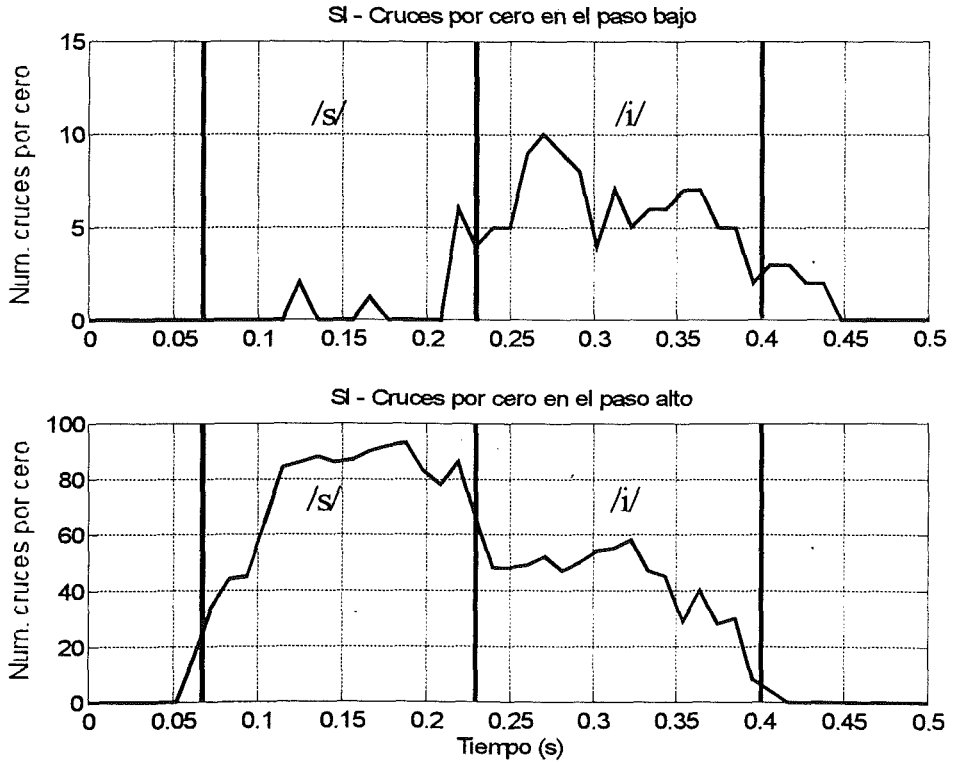

Figura 6.6: Pasos por cero de la señal paso-bajo y paso-alto para el vocablo SI

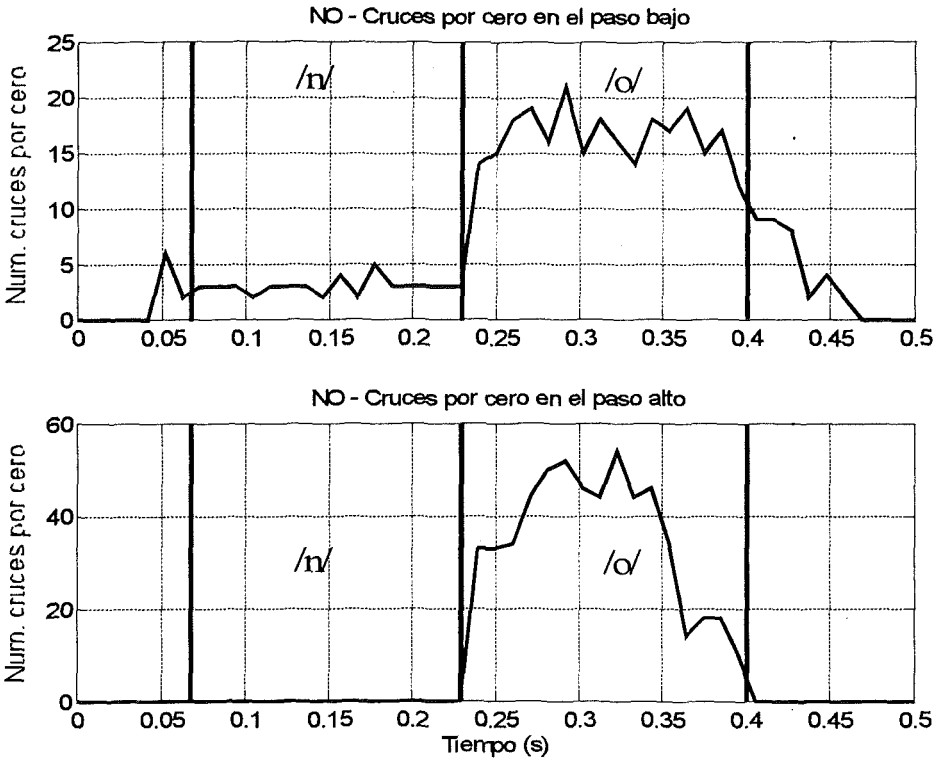

Figura 6.7: Pasos por cero de la señal paso-bajo y paso-alto para el vocablo NO

en el problema, en aras de simplificar la etapa de discriminación temporal difusa posterior. Esta simplificación consiste en la síntesis de un sistema difuso tradicional que suministrándole a la entrada los pasos pos cero de los canales paso-bajo y paso-alto obtenga una salida que represente al fonema que se está produciendo en cada instante. La construcción del citado sistema se presenta en la sección siguiente.

6.5 Clasificación de fonemas

A la vista de las gráficas de la Figura 6.6 y la Figura 6.7 pueden apreciarse claramente las diferencias entre los cuatro fonemas que se han pronunciado (/s/, /i/,/n/ y /ó/). Observamos que la /s/ tiene muchos pasos por cero en la banda paso-alto y casi ninguno en la paso-bajo, pudiendo expresar estos valores como ~90 y ~0, respectivamente. En general, este comportamiento será parecido para los fonemas silbantes y fricativos, como /sh/ o /f/, puesto que son sonidos similares, cuyo contenido espectral se sitúa básicamente en la banda de altas frecuencias. Las vocales /i/ y /o/ tienen ambas bastantes pasos por cero en la banda paso-bajo, aunque es posible distinguir la /i/

de la /o/ por el hecho de que el número de pasos por cero (PPC) de la primera en esta banda es algo inferior. En la banda paso-alto, ambas vocales tienen alrededor de la mitad de PPC que la /s/, unos 50. Por último, la /n/ no tiene PPC en la banda paso-alto y tiene algunos PPC en la paso-bajo (entre 3 y 5). Evidentemente, los silencios (/-/) no presentan actividad en ninguna de las dos bandas. Es de destacar que las diferencias observadas en los cruces por cero se han puesto de manifiesto utilizando conscientemente expresiones difusas, como "muchos" o "bastantes".

Con la información extraída en el parágrafo precedente, puede construirse un sistema difuso que efectúe la clasificación de los pasos por cero en fonemas. Tal sistema puede describirse en base al conjunto de 5 reglas siguientes, en las que "ppcpb" significa "*Número de cruces por cero en el canal paso-bajo en 10 milisegundos*", y "ppcpa" lo mismo para el canal paso-alto:

1. Si ppcpb es cero y ppcpa es cero entonces la salida es /-/

2. Si ppcpb es bajo y ppcpa es cero entonces la salida es /n/

3. Si ppcpb es medio y ppcpa es medio entonces la salida es /i/ (6.1)

4. Si ppcpb es alto y ppcpa es bajo entonces la salida es /ó/

5. Si ppcpa es alto entonces la salida es /s/ (indep. del paso-bajo)

En estas reglas, los términos lingüísticos "bajo", "medio", etc... se representan mediante conjuntos difusos cuya posición y forma exacta pueden determinarse de formas muy diversas, entre las que se encuentran la intuición, los métodos estadísticos, el aprendizaje automático, o la clasificación difusa [Ross 1995], [Klir 1995]. Sin embargo, dada la sencillez del problema, en una primera fase se ha optado por realizar el diseño manualmente, procediendo a optimizarlo en una fase posterior mediante un algoritmo de entrenamiento adecuado.

El clasificador difuso de fonemas será, pues, un sistema difuso con dos entradas (cruces por cero paso alto y paso bajo), y una salida numérica cuyo valor nos indicará qué tipo de fonema está presente en la entrada en cada momento. El siguiente cuadro

muestra la correspondencia entre el valor de la salida y la categoría de fonema reconocido:

Tabla 6-1: Correspondencia entre el fonema reconocido y la salida del clasificador difuso

Fonema	Salida
"Silencio" (/-/)	0
"Nasal" (/n/)	1
"Vocal_1" (/i/)	2
"Vocal_2" (/ó/)	3
"Silbante" (/s/)	4

Para la realización se este sistema se ha escogido una estructura tipo Sugeno, con cinco *singleton* en el espacio de salida ubicados cada uno en un valor entero entre 0 y 4. El espacio de cada una de las dos entradas se particiona en cuatro conjuntos difusos, "cero", "bajo", "medio" y "alto", cuyas funciones de pertenencia pueden verse en la Figura 6.8. La razón de optar por un sistema tipo Sugeno en lugar de uno tipo Mamdani es que los primeros presentan una desfuzzificación mucho más simple, realizándose en este caso mediante una simple media ponderada. Además, los algoritmos de que se dispone para entrenar el sistema sólo aceptan como entrada sistemas Sugeno. De todas formas, esta elección no es crítica pues ha sido ya ampliamente demostrado que ambos tipos de sistemas permiten obtener los mismos resultados.

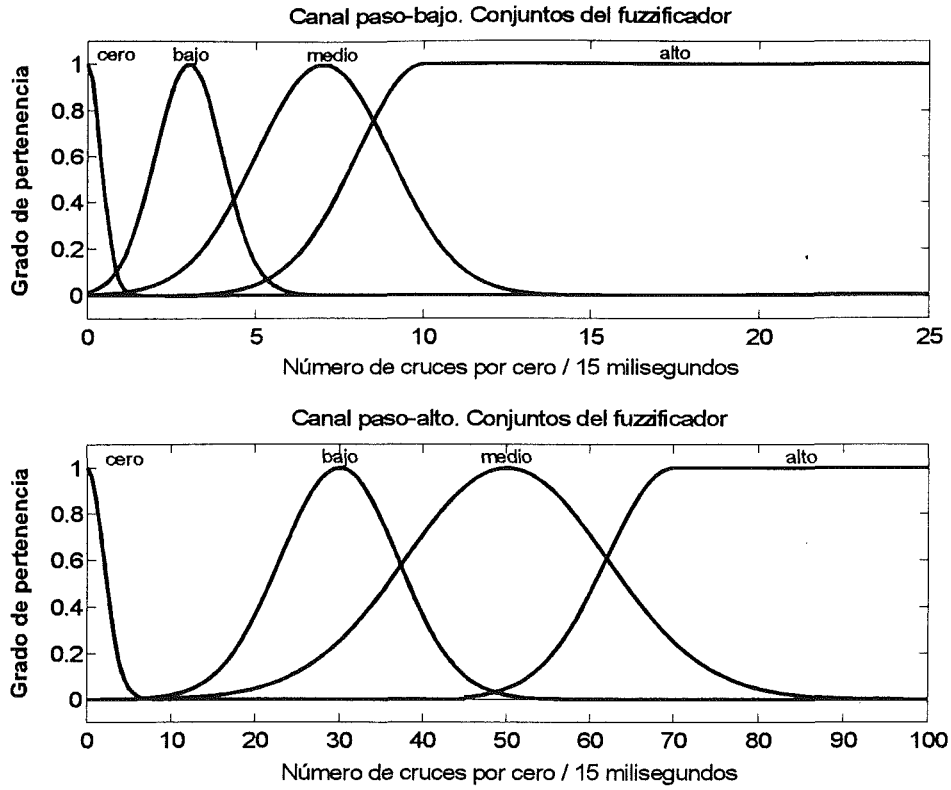

Figura 6.8: Partición difusa de las entradas "ppcpb" y "ppcpa"

Como se ha mencionado anteriormente, en una primera fase los parámetros de estas funciones de pertenencia se han obtenido mediante la observación de multitud gráficas como las de la Figura 6.6 o Figura 6.7, resultado de pronunciar múltiples "si" y "no" por varios locutores. Para comprobar el funcionamiento del clasificador una vez diseñado, se ha procedido a capturar una secuencia de voz en la que se ha pronunciado la sucesión "SI...NO...NO...SI". De esta señal se han calculado los cruces por cero en las bandas paso alto y paso bajo, y éstos se han introducido al clasificador, obteniéndose a su salida el resultado mostrado en la Figura 6.9. En ella puede observarse que el sistema funciona aceptablemente bien, si se considera que el tiempo total invertido hasta ahora en el diseño ha sido mínimo. La señal de salida está básicamente cuantizada a los cinco valores deseados según la combinación presente a las entradas. De todos modos, el funcionamiento del sistema no puede ser perfecto dada la variabilidad de la señal de voz y lo simple de la parametrización. Esto queda patente al observar el fallo de clasificación sufrido en el fonema /i/ del último SI: la salida en este punto debería ser de 2 y en cambio fluctúa entre 2 y 3 puesto que el sistema confunde la /i/ con la /ó/.

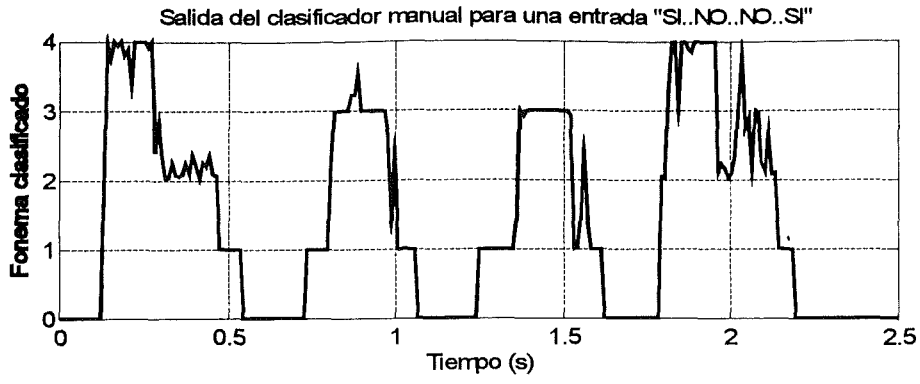

Figura 6.9: Salida del sistema clasificador difuso obtenido manualmente

Otra observación interesante que puede hacerse es que el clasificador suele dar una salida errónea en los tramos finales de cada palabra. Un estudio del problema permite establecer que la causa de ello es que en esos instantes la energía de la señal está decreciendo hasta anularse, sucediendo lo mismo con los cruces por cero puesto que éstos se calculan con algo de histéresis. Debido a ello, para conseguir reconocer correctamente la palabra deberá despreciarse este tramo de señal en las etapas posteriores.

Para tratar de mejorar, si cabe, las características del clasificador obtenido, se ha procedido seguidamente a modificar los conjuntos difusos de las entradas y salidas mediante un algoritmo de optimización adecuado. Para ello pueden utilizarse técnicas clásicas de optimización o bien métodos más modernos como los algoritmos genéticos o el aprendizaje mediante redes neuronales. En esta última categoría existe hoy en día un algoritmo bastante extendido llamado ANFIS [Jang 1993], que además está perfectamente integrado en el paquete de *software* MATLAB utilizado para efectuar las simulaciones que aquí se presentan. Por ello se ha escogido este algoritmo para entrenar el clasificador anterior, para lo cual deben suministrarse al ANFIS un número suficiente de "ejemplos" formados por pares de valores entrada-salida. El resultado del entrenamiento puede observarse en la Figura 6.10, consistente en la salida del clasificador optimizado para la misma entrada que en el caso anterior.

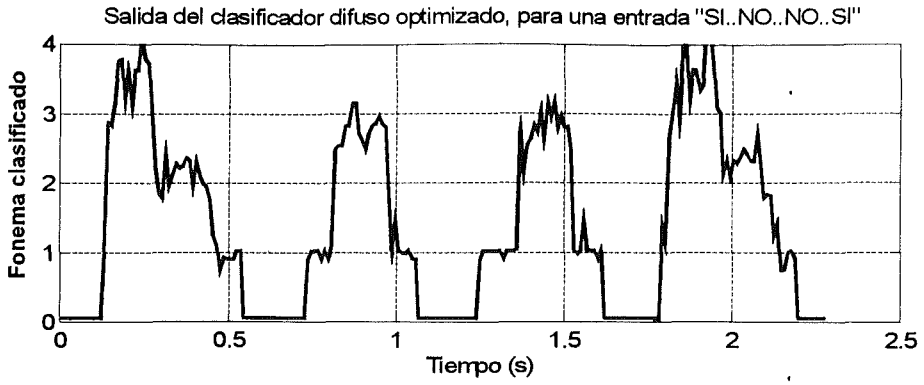

Figura 6.10: Salida del sistema clasificador difuso optimizado

Puede observarse que ahora el clasificador es capaz de distinguir mejor las /i/, a costa quizá de perder capacidad de decisión en las /s/. Esto puede ser indicativo de que el sistema no se ha entrenado con un número suficiente de muestras, o de que se está llegando al límite de discriminación que es posible alcanzar con la parametrización efectuada. Este clasificador presenta además varios problemas adicionales que no se reflejan en la Figura 6.10: En primer lugar la disposición final de los conjuntos difusos resulta mucho menos intuitiva. En segundo lugar el algoritmo de entrenamiento sustituye los conjuntos difusos a la salida del sistema de inferencia, que eran *singleton* constantes en el clasificador manual, por funciones lineales de las entradas; es decir, trabaja con un sistema Sugeno de orden 1 en lugar de uno de orden cero. Ello implica por un lado que su realización es más compleja, y por otro que para ciertas situaciones de entrada no presentes en los valores de entrenamiento pueden darse salidas fuera del rango [0...4], cosa que puede constituir también un problema de cara a la imlpementación práctica del sistema.

Sea como sea, ya que los resultados de ambos han resultado ser similares, se ha optado en este caso por trabajar con el clasificador manual, más simple e intuitivo, sometiéndolo a un breve refinado para conseguir que distinga algo mejor las /i/, con lo que queda concluida la etapa de diseño del clasificador.

6.6 *Discriminador basado en conceptos temporales difusos*

Hasta ahora el sistema se limita a clasificar cada fonema en una de las cuatro categorías definidas. Para proseguir en el proceso de discriminación entre el "SI" y el "NO" es preciso además tener en cuenta la *posición relativa* en que se producen los

distintos fonemas, así como su *duración*. Al considerar estas variables deben introducirse conceptos temporales en los razonamientos. Por ende, dichos conceptos temporales no pueden ser demasiado precisos, pues el instante de emisión y la duración de los distintos fonemas variará entre dos producciones del mismo vocablo pronunciadas en instantes distintos y/o por distintos locutores. Además, debido a la variabilidad intrínseca de la señal de voz, la salida del clasificador no es siempre 0, 1, 2, 3 o 4 *exactamente*, como ya puede apreciarse en la Figura 6.9 y la Figura 6.10. Cabe pues considerar esta salida como una variable imprecisa y tratarla también de forma difusa.

Por todo ello puede pensarse en diseñar otro sistema difuso que discrimine entre un "SI" y un "NO" a partir de la señal suministrada por el clasificador anterior, a la que llamaremos x, y el conjunto de reglas siguiente:

- **Si** (x ANTES era /s/) **y** (x AHORA es /i/) **entonces** (palabra es SÍ)
- **Si** (x ANTES era /n/) **y** (x AHORA es /ó/) **entonces** (palabra es NO)

en las que las expresiones /s/, /i/, /n/ y /ó/ se consideran ahora números difusos centrados en 4, 2, 1 y 3, respectivamente. En esta situación resulta de evidente utilidad la formulación propuesta en el capítulo 4 de esta tesis para realizar la inferencia difusa en presencia de los conceptos temporales "antes" y "ahora". Es de destacar que en el tipo de problema que aquí se trata deben considerarse los citados conceptos temporales como *inclusivos*, puesto que los fonemas tienen *duración*. Con ello se evita que el sistema responda con salidas erróneas debido a la presencia de situaciones espúreas en algún punto de la cadena de reconocimiento, las cuales pueden ser interpretadas accidentalmente como un fonema válido. De este modo, el algoritmo que deben utilizarse para realizar la inferencia temporal difusa es la convolución lineal analógica dada por la ecuación (4.28), que adecuadamente particularizada para este caso permite obtener los grados de verdad de cada proposición de las reglas mediante

$$w_1(t) = \mathrm{T}(x \text{ Antes era } /\mathrm{s}/) = \int_{-\infty}^{t} \tilde{\mu}_{\text{Antes}}(t-\tau) \cdot \mu_{/s/}(x(\tau)) \cdot d\tau \qquad (6.2)$$

$$w_2(t) = \mathrm{T}(x \text{ Ahora es } /\mathrm{i}/) = \int_{-\infty}^{t} \tilde{\mu}_{\text{Ahora}}(t-\tau) \cdot \mu_{/i/}(x(\tau)) \cdot d\tau \qquad (6.3)$$

$$w_3(t) = \mathrm{T}(x \text{ Antes era } /\mathrm{n}/) = \int_{-\infty}^{t} \tilde{\mu}_{\text{Antes}}(t-\tau) \cdot \mu_{/n/}(x(\tau)) \cdot d\tau \qquad (6.4)$$

$$w_4(t) = \mathrm{T}(x \text{ Ahora es } /\mathrm{o}/) = \int_{-\infty}^{t} \tilde{\mu}_{\text{Ahora}}(t-\tau) \cdot \mu_{/o/}(x(\tau)) \cdot d\tau \qquad (6.5)$$

Cada uno de estos grados de verdad se obtiene, pues, como salida de un circuito lineal cuya respuesta al impulso sea la función $\tilde{\mu}_{Antes}(t)$ o $\tilde{\mu}_{Ahora}(t)$ y cuya entrada sea la salida de un fuzzificador que devuelva el grado de pertenencia de la salida del clasificador al concepto difuso /s/, /i/, /n/ o /ó/, respectivamente. A modo de ejemplo, en la Figura 6.11 se presenta el circuito que permite encontrar el grado de verdad $w_1(t)$:

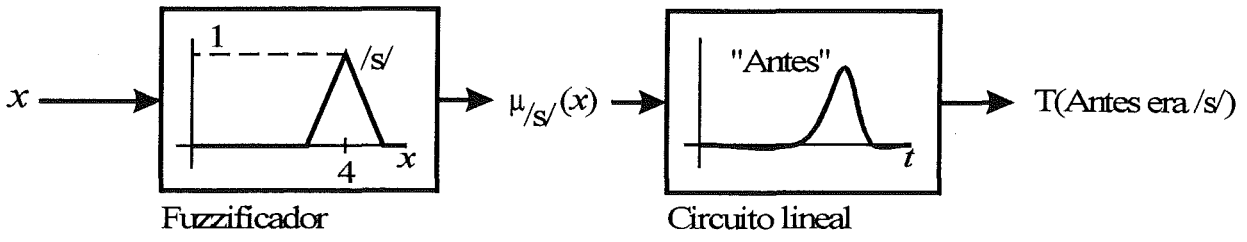

Figura 6.11: Circuito para la realización de la ecuación (6.2)

Después de cada uno de estos circuitos, el grado de verdad de un antecedente completo se obtendrá haciendo el Y lógico de los grados de verdad de las dos premisas que lo componen, para lo cual puede utilizarse el mínimo o el producto, aunque en el prototipo construido se ha utilizado el primero por tener una implementación analógica más sencilla.

Las expresiones concretas para las funciones de pertenencia $\tilde{\mu}_{Antes}(t)$ o $\tilde{\mu}_{Ahora}(t)$ deben obtenerse en este caso mediante las técnicas expuestas con anterioridad en el capítulo 5, pues se está trabajando en el dominio analógico. Para ello debe especificarse primero unas plantillas para los conceptos temporales "Antes" y "Ahora" y encontrar las funciones de red del circuito que mejor las aproximan, establecer un compromiso aceptable entre el grado de parecido deseado y el orden del circuito resultante. La observación cuidadosa de la salida del clasificador difuso permite deducir que la duración promedio de los fonemas es de unos 150 milisegundos, y que están situados consecutivamente, sin intervalo apreciable entre ellos. También se aprecia que a la salida del clasificador las /n/ tienen una duración algo menor, de unos 100 milisegundos, algo que no se había detectado en una primera fase. En estas duraciones están incluidos

también los transitorios de ataque y relajación de cada fonema. Si los despreciamos, podemos fijar la duración de la parte central del fonema entre 50 y 100 milisegundos. Como información adicional cabe tener en cuenta que la duración de la respuesta al impulso de un circuito determina la rapidez con que éste responde a cambios en la señal de entrada, por lo que una respuesta impulsional muy corta proporcionará más velocidad de reacción frente a un patrón válido, pero a su vez también hará el sistema más sensible a clasificaciones erróneas espúreas.

Con esta información, unas posibles plantillas para los conceptos temporales difusos "Antes" y "Ahora" se muestran en la Figura 6.12.

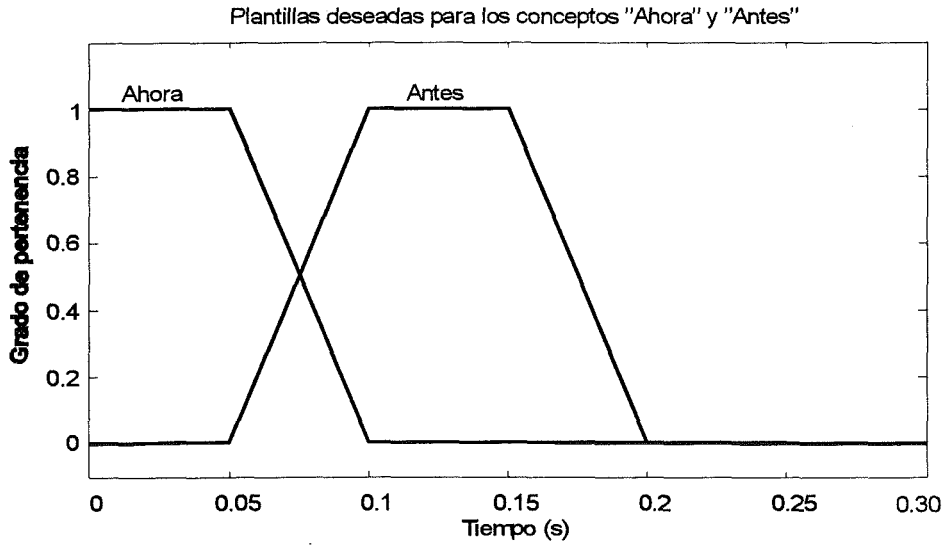

Figura 6.12: Plantillas para las respuestas al impulso de los circuitos que sintetizan los conceptos "Antes" y "Ahora"

Llegados a este punto es obvio que lo único que resta decidir es el orden de los circuitos que van a utilizarse para aproximar estas plantillas. Para apoyar esta decisión se propone a continuación valorar los resultados obtenidos en dos casos diferenciados: el primero primando la sencillez y el segundo la exactitud. En el primer caso se han utilizado para sintetizar el "Ahora" y el "Antes" circuitos de primer y segundo orden, respectivamente. Como puede observarse en la Figura 6.13, en este caso las aproximaciones son mas bien groseras, especialmente la de primer orden. Sin embargo son muy sencillas de construir y de ajustar.

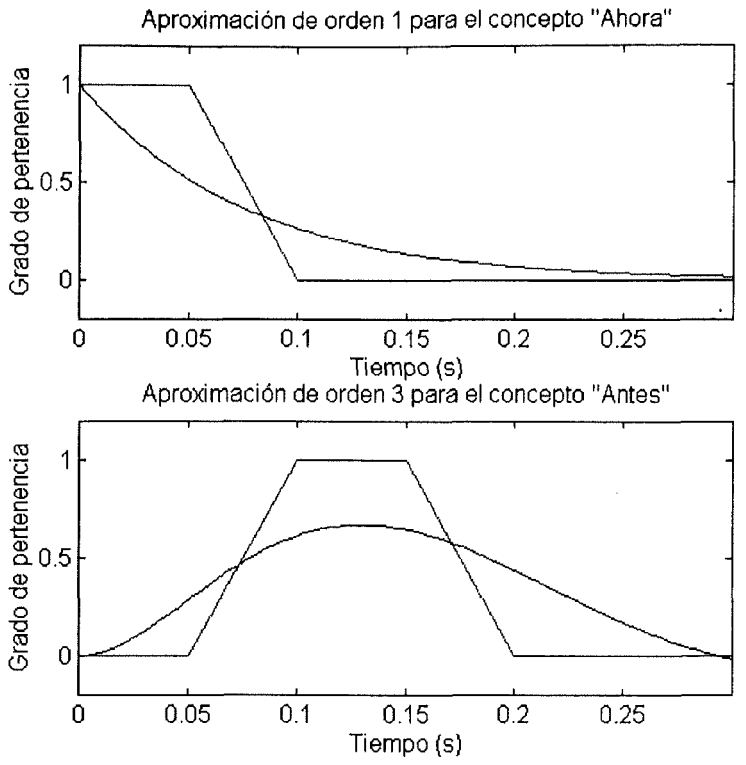

Figura 6.13: Aproximaciones de primer y segundo orden para el "Ahora" y el "Antes"

Concretamente, las funciones de transferencia de estos circuitos, **aún sin normalizar a área unidad**, son:

$$H_{\text{Ahora}}(s) = \frac{1}{s + 13.33}$$

$$H_{\text{Antes}}(s) = \frac{384}{s^3 + 28s^2 + 480s + 3840}$$

cuya realización se obtiene mediante un circuito RC simple en el primer caso, y en el segundo mediante un circuito activo con un amplificador operacional.

Para obtener una aproximación más exacta a los conceptos temporales deseados, se ha probado también con unos circuitos de orden cuarto. La Figura 6.14 muestra la respuesta impulsional de estos circuitos, que como puede observarse se ajusta mucho más a las plantillas propuestas, siendo su mayor inconveniente la mayor complejidad de su realización física.

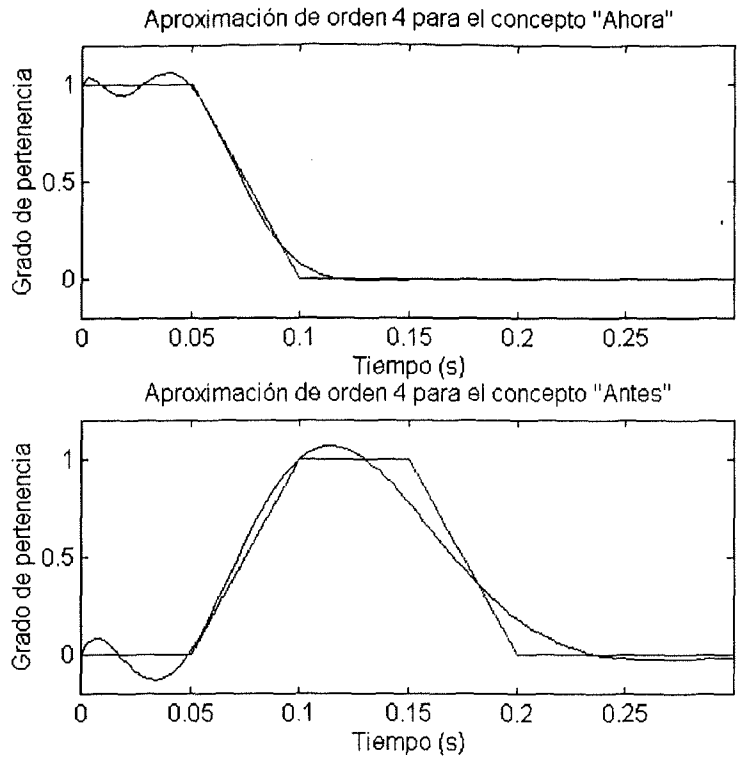

Figura 6.14: Aproximaciones de orden cuarto para el "Ahora" y el "Antes"

En este segundo caso las funciones de transferencia obtenidas resultan engorrosas de transcribir, por lo que en su lugar se da en la Tabla 6-2 un listado de los parámetros de cada circuito en formato Z-P-K.

Tabla 6-2: Parámetros de los circuitos "Ahora" y "Antes"

	z (ceros)	*p* (polos)	*k*
Circuito "Ahora"	-201.31 -15.91 + 83.72j -15.91 - 83.72j	-44.05 +55.91j -44.05 -55.91j -59.35 +18.04j -59.35 -18.04j	**1**
Circuito "Antes"	52.186 + 31.43j 52.186 - 31.43j	-20.30 + 28.70j -20.30 - 28.70j -28.25 + 9.19j -28.25 - 9.19j	**29.39**

Estos parámetros permiten expresar la función de transferencia como

$$H(s) = k \cdot \frac{\prod_{j=1}^{m}(s - z_j)}{\prod_{i=1}^{n}(s - p_i)}$$

6.7 Resultados

Habiendo encontrado los circuitos que realizan las aproximaciones de los conceptos temporales difusos deseados, resta sólo realizar circuitalmente cada proposición de la regla utilizando una estructura como la mostrada en la Figura 6.11. Dado que cada regla tiene dos proposiciones en su antecedente, habrá que sintetizar dos fuzzificadores y conceptos temporales para implementarla, y efectuar el *and* lógico de los grados de verdad de cada proposición para obtener el grado de verdad global de la regla. Posteriormente deberá utilizarse este último valor para afectar al conjunto consecuente de la regla en cuestión.

Nótese, sin embargo, que esta última operación no es necesaria en este caso, pues al fin y al cabo lo que aquí se pretende es decidir si se ha pronunciado un SI o un NO, para lo cual basta un comparador. De este modo el sistema final resulta más simple si se toman como salida los grados de verdad de cada una de las dos reglas, que son de hecho dos tensiones que oscilan entre los valores 0 y 1. Si una de las dos supera un umbral predefinido, por ejemplo 0.7, se habrá producido el reconocimiento de una de las dos palabras.

La Figura 6.15 muestra estas dos curvas para el caso de utilizar los conceptos temporales de orden bajo mostrados en la Figura 6.13. La frase que se ha pronunciado delante del micrófono es en este caso la misma que se mostraba en el clasificador, es decir, la secuencia "SI...NO...NO...SI". Como puede observarse, el grado de verdad del "SI", en color verde, presenta una subida justamente después de haberse producido este vocablo. De forma parecida, el grado de verdad del "NO", en rojo, presenta un máximo después de haberse dado un sonido nasal seguido de uno vocálico. Ambos máximos tienen la suficiente amplitud como para permitir su detección y, sobre todo, su discriminación, aunque el grado de verdad máximo al que llegan no es muy elevado,

siendo inferior a 0.5 en algún caso. Ello puede comportar que algunos SI o NO no sean reconocidos por no llegar al nivel mínimo que se fije.

Como muestra la Figura 6.16, esta situación mejora sensiblemente mediante la utilización de unos circuitos de mayor orden en la realización de los conceptos temporales difusos. En ella se observan los grados de verdad de las dos reglas para la misma entrada que en el caso anterior pero utilizando ahora los circuitos de cuarto orden cuyas respuestas al impulso se mostraban en la Figura 6.14.

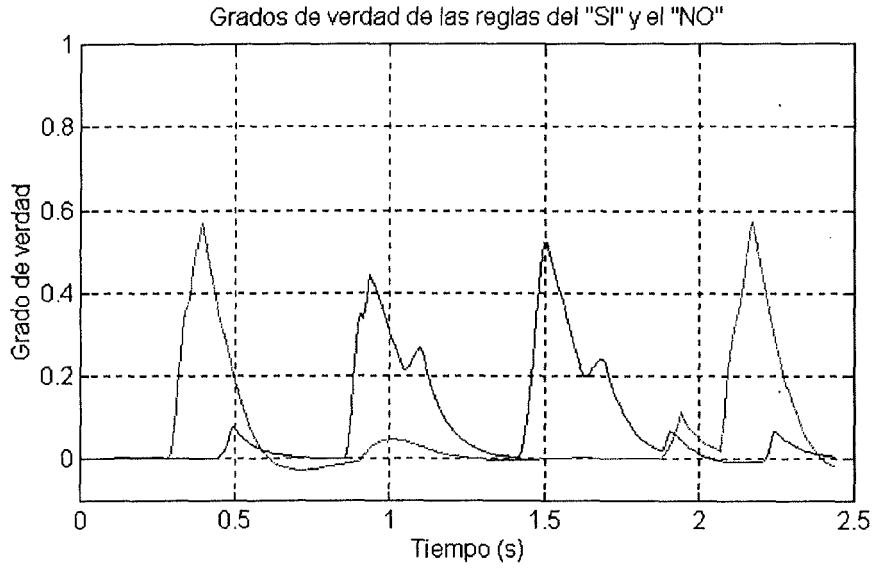

Figura 6.15: Salida del discriminador para los conceptos temporales de la Figura 6.13

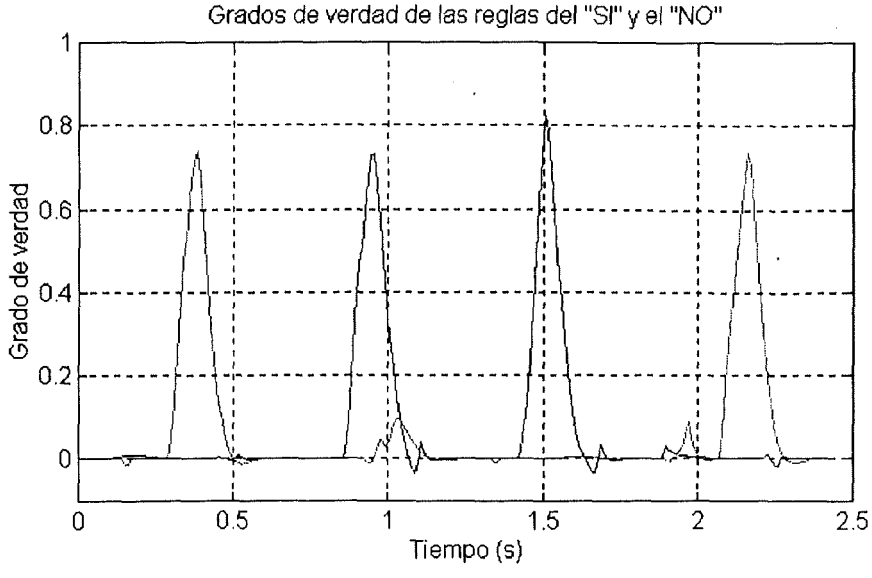

Figura 6.16: Salida del discriminador para los conceptos temporales de la Figura 6.14

Con estos nuevos conceptos temporales se consigue en el mejor caso elevar el valor de la salida en más del 85%, acusándose el cambio con mayor intensidad en los NO.

En la implementación real se han hecho varias simplificaciones. Una de ellas es diseñar el primer clasificador de tal forma que no distinga entre la /i/ y la /ó/, sino que etiquete a ambas como "vocal". De este modo se simplifica tanto el diseño del clasificador como el de la ultima etapa, puesto que hay que sintetizar un fuzzificador y un concepto temporal menos. Además, los valores de la salida resultan ser también algo más elevados, a costa de no distinguir, por ejemplo, un SI de un SO. Otras simplificaciones afectan a la forma de las funciones de pertenencia de los fuzzificadores, que se han tomado lineales a tramos por ser de más fácil realización mediante amplificadores operacionales. Las señales de salida se han utilizado para encender unos diodos LED de color verde y rojo en respuesta a la pronunciación de un SI o un NO, respectivamente.

Las gráficas siguientes corresponden ya a señales presentes en el circuito. En ellas puede observarse la salida del clasificador cuando se pronuncia "SI...NO".

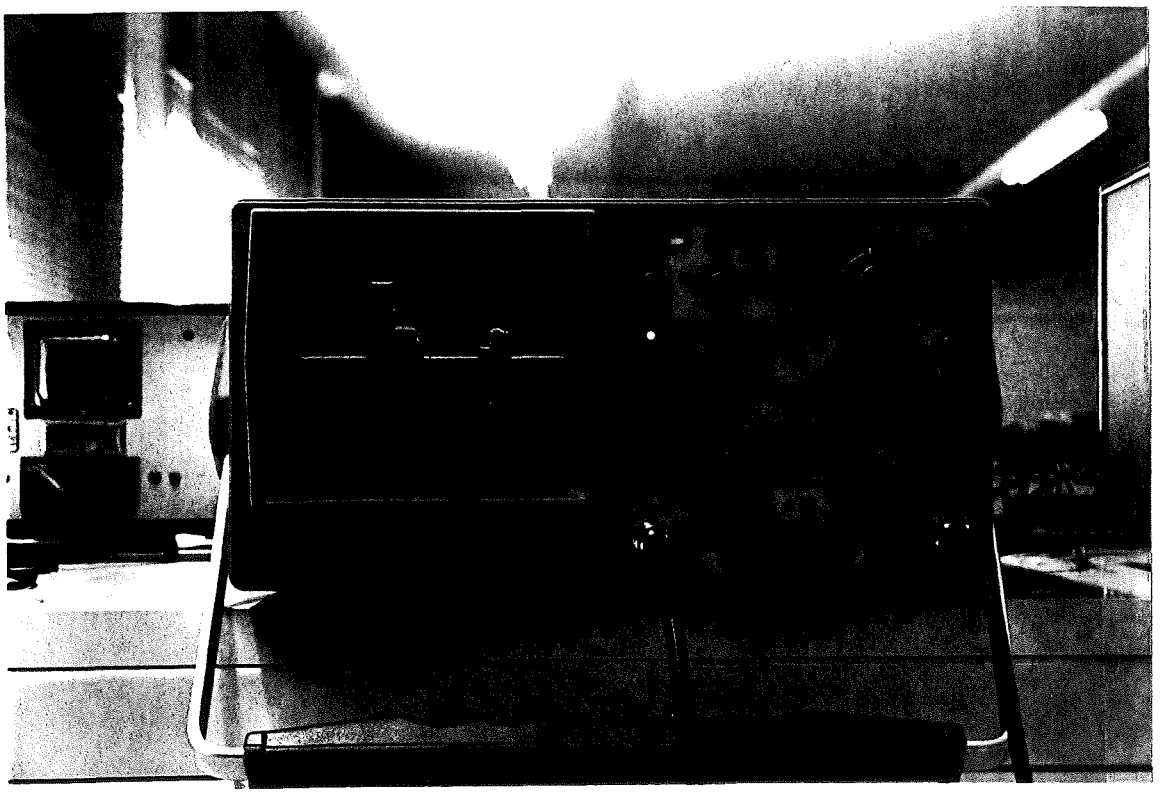

Figura 6.17: Salida del clasificador en el circuito real

Figura 6.18: Discriminador de voz. Visión de conjunto.

Como se ha puesto de manifiesto, el uso de técnicas temporales difusas permite discriminar fenómenos que se producen en instantes distintos, pero cuya duración y localización (además de su amplitud) son imprecisas. Dichas técnicas se han aplicado con éxito a la construcción de un prototipo capaz de discriminar entre los vocablos "SI" y "NO" usando íntegramente tecnología analógica, lo que lo convierte en un dispositivo muy económico y de reducidas dimensiones, que admite una fácil integración si se toman como base sistemas de inferencia difusa analógicos existentes actualmente. En este sentido, el presente prototipo, desarrollado inicialmente con componentes discretos, se prevé integrarlo usando el *chip* analógico de inferencia difusa diseñado por Vidal y Rodríguez-Vázquez [Vidal 1994], [Vidal 1995]. Cabe destacar que en un caso como el presente la solución analógica puede ser de interés frente a la digital, por tres razones básicas:

1. Es más rápida que cualquier microcontrolador de propósito general ejecutando un programa que realice en tiempo real todas las operaciones que se han descrito. En particular, el muestreo y el

filtrado posterior también deberían hacerse probablemente de forma analógica en el segundo caso.

2. La complejidad de integración es menor en el caso analógico. Se podría decir que la solución analógica permite obtener mayor "inteligencia" por transistor integrado que la digital, siendo por lo tanto más rentable.

3. La compatibilidad electromagnética de los circuitos analógicos es mayor que la de los digitales. Los sistemas digitales generan bastante ruido, el cual puede afectar seriamente el sistema de reconocimiento pues la parametrización escogida (cruces por cero) es muy sensible a señales interferentes, aunque tengan poca amplitud.

Por otra parte, se ha comprobado que la parametrización de la señal vocal mediante los cruces por cero en las bandas paso-bajo y paso-alto de la señal es más que suficiente para los objetivos prefijados. Esto sugiere la posibilidad de ampliar el número de palabras a reconocer manteniendo básicamente la misma parametrización. En este sentido, se tiene proyectada la construcción de un prototipo capaz de reconocer los dígitos del cero al nueve, con independencia del locutor. También se puede pensar en usar una parametrización menos tosca de la señal vocal para aumentar la potencia de reconocimiento del dispositivo.

Como nota final cabe comentar que el potencial campo de aplicación de esta técnica es amplio, y que a juzgar por su económica implementación sus aplicaciones más inmediatas podrían centrarse en juguetería (control por voz económico), telefonía (marcaje por voz) y, en general, cualquier aplicación en donde se requiera discriminar entre un número reducido de vocablos sin tener que recurrir a las técnicas clásicas de procesado de señal, más precisas pero también más costosas.