COMPARACIÓN ENTRE ANÁLISIS WAVELETS Y FOURIER APLICADOS AL RECONOCIMIENTO AUTOMÁTICO DEL HABLA

TESIS QUE PRESENTA

HUGO LEONARDO RUFINER DI PERSIA

PARA LA OBTENCIÓN DEL GRADO DE

MAESTRO EN INGENIERÍA BIOMÉDICA

ASESOR

DR. JOHN GODDARD CLOSE

DICIEMBRE DE 1996

Casa Abierta al Tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA - IZTAPALAPA
DIVISIÓN CIENCIAS BÁSICAS E INGENIERIA

A mi madre Silvia Di Persia,

Por su ejemplo

A mi esposa Stella,

Por su apoyo y comprensión

A mi hijo Juan,

Por su ternura y alegría

Agradecimientos

- A la Universidad Nacional de Entre Ríos (UNER), a la Provincia de Entre Ríos, y a la Organización de Estados Americanos (OEA) por el soporte económico que permitió mi estancia en México y la realización de esta tesis.

- A mi asesor, el Dr. John Goddard, por su constante disposición al debate y su valiosa orientación sin la cuál seguramente este trabajo no hubiera sido el mismo.

- Al Ing. Agustín Carpio y a la Med. Susana Perrone por su intervención en el Convenio de Intercambio de Recursos Humanos (UNER-UAMI) que posibilitó mis estudios.

- Al M. Miguel Cadena Mendez por su gestión que hizo posible esta experiencia.

- A mis compañeros del Laboratorio de Audiología de la UAMI, por favorecer el clima ameno de trabajo y convivencia que posibilitó mi labor.

- A todos mis compañeros de la FI-UNER por su apoyo constante a pesar de distancias y contratiempos.

Prefacio

La emulación de la forma de comunicación humana por las computadoras ha sido una meta largamente perseguida. Alcanzarla permitiría interactuar con nuestras máquinas de una manera más sencilla y completamente distinta a la actual. Durante el desarrollo de mi tesis de licenciatura abordé una parte de la solución de este vasto problema a través de diversas técnicas de Inteligencia Artificial. Posteriormente encaré el análisis de la voz mediante técnicas clásicas y modelos de oído. Aquí comenzó a aparecer la idea de comparar un análisis similar al que realiza nuestro oído (como el análisis Wavelets) con las técnicas clásicas basadas en estimadores espectrales (como el análisis de Fourier).

En el presente trabajo se pretende evaluar la mejor de estas dos alternativas (Wavelets y Fourier) para la etapa de preprocesamiento de un Sistema de Reconocimiento Automático del Habla utilizando Redes Neuronales.

Esta tesis se organizará de la siguiente manera: En la primera parte se introducirá al lector en el problema, terminología y antecedentes. En el capítulo siguiente se abordarán someramente los aspectos fisiológicos más relevantes de la comunicación humana para señalar los mecanismos responsables de la producción, emisión, recepción y reconocimiento del habla. El comprender la naturaleza de la señal de voz nos permitirá comprender que parámetros son más importantes para la discriminación de los distintos fonemas y el conocer como funciona el sistema auditivo nos autoriza a evaluar otros tipos de análisis. En el capítulo tercero se describen los datos utilizados en los experimentos y los criterios para su recolección y elección. El capítulo siguiente se ocupa del tema del procesamiento de la señal de voz. Se introduce así el análisis clásico y el basado en Wavelets, describiendo las principales familias de Wavelets y presentando un método para escoger el análisis que mejor discrimine entre los fonemas elegidos. En el capítulo cinco se expone lo referente a las redes neuronales que actúan como clasificadores de los patrones generados en la etapa de análisis en los distintos fonemas. Aquí se describen las arquitecturas de redes que permiten aprovechar de manera eficiente los aspectos dinámicos de la señal de voz para su clasificación o reconocimiento. En el capítulo seis se presentan los resultados de los experimentos junto con su interpretación y conclusiones. Finalmente se presentan las referencias y la bibliografía empleada durante el trabajo.

Indice

I . INTRODUCCIÓN 1

ANTECEDENTES Y DEFINICIONES 1
DESCRIPCIÓN DEL TRABAJO 6

II . ASPECTOS FISIOLÓGICOS 8

INTRODUCCIÓN 8
MECANISMO DE PRODUCCIÓN DEL HABLA 8
LA SEÑAL DE VOZ 13
FISIOLOGÍA DE LA AUDICIÓN 15

III . LOS DATOS 22

INTRODUCCIÓN 22
DESCRIPCIÓN DE TIMIT 23
Organización de los datos 24
Tipos de Archivo 24
Selección de Hablantes 25
Condiciones de Grabación 26
Texto del Corpus 26
Subdivisión en Entrenamiento y Prueba 27
Códigos de Símbolos Fonémicos y Fonéticos 28
DATOS ELEGIDOS PARA LOS EXPERIMENTOS 30

IV . EL PROCESAMIENTO 37

INTRODUCCIÓN 37
TRANSFORMADA DE FOURIER 38
Transformada de Fourier de Tiempo Corto 38
TRANSFORMADA WAVELET 40
FUNDAMENTOS TEÓRICOS Y DEFINICIONES 45
Propiedades del Análisis Multiresolución 49
TRANSFORMADA WAVELET DISCRETA 49
FAMILIAS DE WAVELETS 50
Meyer 52
Daubechies 54
Symmlets 54
Coiflets 57
Splines 57
Vaidyanathan 62
ELECCIÓN DE LA BASE ÓPTIMA 62
ASPECTOS DE IMPLEMENTACION PRÁCTICA 64

V . EL CLASIFICADOR 66

INTRODUCCIÓN 66
REDES NEURONALES ESTATICAS : PERCEPTRON MULTICAPA 68
REDES NEURONALES DINAMICAS 70
Extensión de Retropropagación para Aprendizaje Temporal 70
Redes Neuronales con retardos temporales 71
Redes de Jordan y Elman 71
Retropropagación a través del tiempo 72
CRITERIOS PARA LA ELECCIÓN DE LA ARQUITECTURA NEURONAL 75
REDES DE KOHONEN 75
ELECCIÓN DE LA FAMILIA DE WAVELETS 76
ASPECTOS DE IMPLEMENTACIÓN PRÁCTICA 78

VI . RESULTADOS Y CONCLUSIONES 82

INTRODUCCIÓN 82
EXPERIMENTOS REALIZADOS 82
RESULTADOS 84
INTERPRETACIÓN Y CONCLUSIONES 89
RECOMENDACIONES Y SUGERENCIAS FINALES 91

VII . REFERENCIAS 92

I . Introducción

Antecedentes y Definiciones

El *Reconocimiento Automático del Habla* (RAH) es un campo multidisciplinario con especial vinculación *al Reconocimiento de Formas* y a *la Inteligencia Artificial* (IA). Su objetivo es la concepción e implementación de sistemas automáticos capaces de interpretar la señal vocal humana en términos de categorías lingüísticas de un universo dado. Según el tipo de categoría, universo y locutor/es presenta distintos grados de complejidad. Para una revisión completa del tema ver [LHR90].

Varias décadas de Investigación y Desarrollo fueron estableciendo la importancia de las siguientes dimensiones en la comprensión de las propiedades de un sistema de reconocimiento automático del habla [LHR90]:

- Dependencia Vs. Independencia del Hablante: un sistema *Dependiente del Hablante* (DH) está entrenado para reconocer solamente una única voz. Un sistema *Independiente del Hablante* (IH) puede reconocer el habla emitida por virtualmente cualquier persona, aunque con menos exactitud.

- Palabras Aisladas Vs. Discurso Continuo: Un sistema de *Reconocimiento de Palabras Aisladas* (RPA) requiere que se efectúen pausas entre las palabras pronunciadas. El *Reconocimiento del Discurso Continuo* (RDC) permite emitir el habla en una forma más natural, pero es más complejo y más sujeto a errores.

- Amplitud del Vocabulario y Complejidad de la Gramática: el vocabulario de un sistema define el conjunto de palabras reconocibles, y la gramática define el tipo de oraciones -o secuencias de palabras- permitidas. Los sistemas de reconocimiento del habla con vocabularios pequeños y gramáticas restrictivas son más fáciles de implementar, pero los sistemas con vocabularios amplios y gramáticas permisivas son más útiles.

- Reconocimiento del Habla Vs. Comprensión del Habla: Un sistema de reconocimiento del habla produce una secuencia de palabras mientras que un sistema de comprensión intenta interpretar la intención del hablante.

Todos los sistemas con interfaces orales -tanto los comerciales como los prototipos experimentales- se esfuerzan en lograr una gran precisión. Generalmente esa meta se alcanza a costa de sacrificar una o más de las dimensiones anteriores. La estructura general de uno de estos sistemas tiene esencialmente tres componentes o etapas

1. Procesamiento o Análisis del Habla: en esta etapa se realiza algún tipo de análisis de la señal de voz en términos de la evolución temporal de parámetros espectrales (previa conversión A/D de la señal). Esto tiene por función hacer más evidentes las

características necesarias para la etapa siguiente. A veces también tiene por objeto reducir la dimensión de los patrones para facilitar también su clasificación.

2. Reconocimiento o Clasificación de Unidades Fonéticas : esta etapa clasifica o identifica los segmentos de voz ya procesados con símbolos fonéticos (fonemas, difonos o sílabas). A veces se puede asociar una probabilidad con este símbolo fonético, lo que permite ampliar la información presentada al siguiente módulo.

3. Análisis en Función de Reglas del Lenguaje : En esta última etapa se pueden aprovechar las reglas utilizadas en la codificación del mensaje contenido en la señal para mejorar el desempeño del sistema y producir una transcripción adecuada. Aquí se utilizan otras fuentes de conocimiento como la Ortografía, la Sintáctica, la Semántica o la Pragmática (ver más adelante).

A veces se agrega un nivel más o que podría llamarse etapa de Comprensión, esto es cuando el sistema está orientado a realizar alguna acción en función de órdenes habladas. Debe aclararse que la separación en reconocimiento y comprensión es arbitraria ya que en realidad nosotros realizamos estos procesos al mismo tiempo y con una fuerte interacción entre ellos. El RAH presenta inconvenientes especiales involucrados con el proceso de comprensión que hacen que un sistema práctico de este tipo sea difícil de implementar [Whi90]. La conversión sin ninguna restricción de la señal analógica en su correspondiente representación fonológica (reconocimiento) no se ha logrado hasta el momento. El lenguaje humano tiene una inmensa complejidad sintáctica y semántica, no comparable con el de ninguna otra especie, y esto nos da una idea del potencial informático necesario para intentar comprenderlo o al menos transcribirlo.

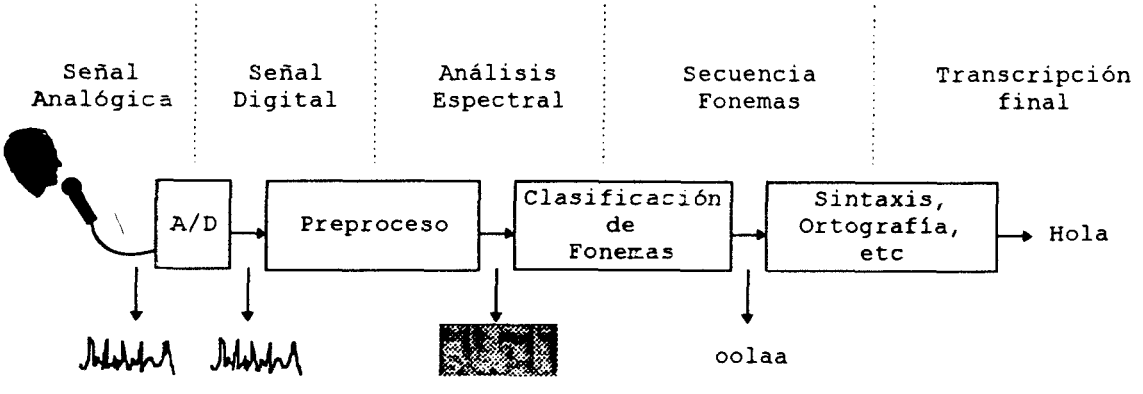

Figura 1: Componentes de un Sistema de Reconocimiento.

Algunos de los problemas se presentan en esta aplicación -especialmente en el RDC- se detallan a continuación [LHR90]:

- Ambigüedad: Existen ambigüedades tanto a nivel de las unidades básicas, como lingüístico (palabras que solo pueden ser interpretadas en el contexto) y también debidas a la coarticulación. Existen palabras con distinto significado y función que poseen la misma representación fonológica (*homófonos*).

- Ruido de fondo: Se produce superposición de la señal de voz pura con el sonido circundante e inclusive con otras voces presentes en el recinto a las cuales no debe *prestarse atención*.

- Variaciones entre hablantes: Existen modificaciones debidas a acentos regionales, dialectos y a las diferentes características propias de cada hablante.

- Variaciones del mismo hablante: Hay variaciones en la señal del mismo hablante, en distintos momentos, debido a distintos estados anímicos como felicidad, depresión, emoción o inclusive debido a cambios en el estado de salud.

- Segmentación: En forma contraria a lo que se podría esperar, la mayoría de las palabras no aparecen separadas en la señal normal -*nosotroslasentendemoseparadas*-, y tampoco es posible separarlas sólo con la información contenida en dicha señal. Esto es debido a la interarticulación entre los fonemas de las palabras adyacentes, su falta de silencio intermedio y además porque algunos fonemas que deben repetirse en su representación escrita aparecen sólo una vez en la señal vocal.

Una idea aproximada de la complejidad involucrada en el reconocimiento de la señal de voz se puede obtener al apreciar un espectrograma de una emisión típica y su correspondiente representación temporal (Figura 2). Este revela las complejas relaciones temporales-frecuenciales que deben establecerse para reconocer o clasificar los fonemas incluidos en la señal, sin tener en cuenta las ambigüedades de orden superior y otras dificultades intrínsecas.

En un sistema típico el proceso de reconocimiento y comprensión del lenguaje hablado consiste en una serie de transformaciones que se aplican a la señal de voz original, estas transformaciones pueden ser vistas como la interpretación de esa señal a la luz de diferentes clases de conocimiento.

La resolución completa del problema del reconocimiento del discurso continuo debe hacer uso de las siguientes fuentes de conocimiento o niveles de análisis [LHR90]:

1. *Fonético:* se encarga de la representación de las características físicas de los sonidos utilizados para la producción del habla.

2. *Fonémico:* se ocupa de la descripción de las variaciones en la pronunciación que aparecen dentro de una palabra o cuando las palabras son dichas juntas en una frase (coarticulación, fusión de sílabas, etc.).

3. *Morfémico:* realiza una descripción del modo en que los morfemas (unidades de significación) son combinados para formar palabras. (formación de plurales, conjugación de verbos, etc.).

4. *Prosódico:* consiste en una descripción de la fluctuación en la acentuación y entonación durante el transcurso de una frase (que también lleva información acerca de lo que se está diciendo).

5. *Sintáctico:* constituye la gramática o reglas de formación de frases, dando lugar a una limitación del numero de frases (no todas las combinaciones de palabras son frases autorizadas).

6. *Semántico:* consiste en analizar el significado de las palabras y las frases que puede ser visto también como una restricción sobre el alcance del mensaje. (no todas las frases validas gramaticalmente tienen significado).

7. *Pragmático:* se ocupa de las reglas de conversación (en un dialogo la respuesta de un interlocutor no debe ser solamente una frase con significado sino también una respuesta razonable a cerca de lo que se esta diciendo).

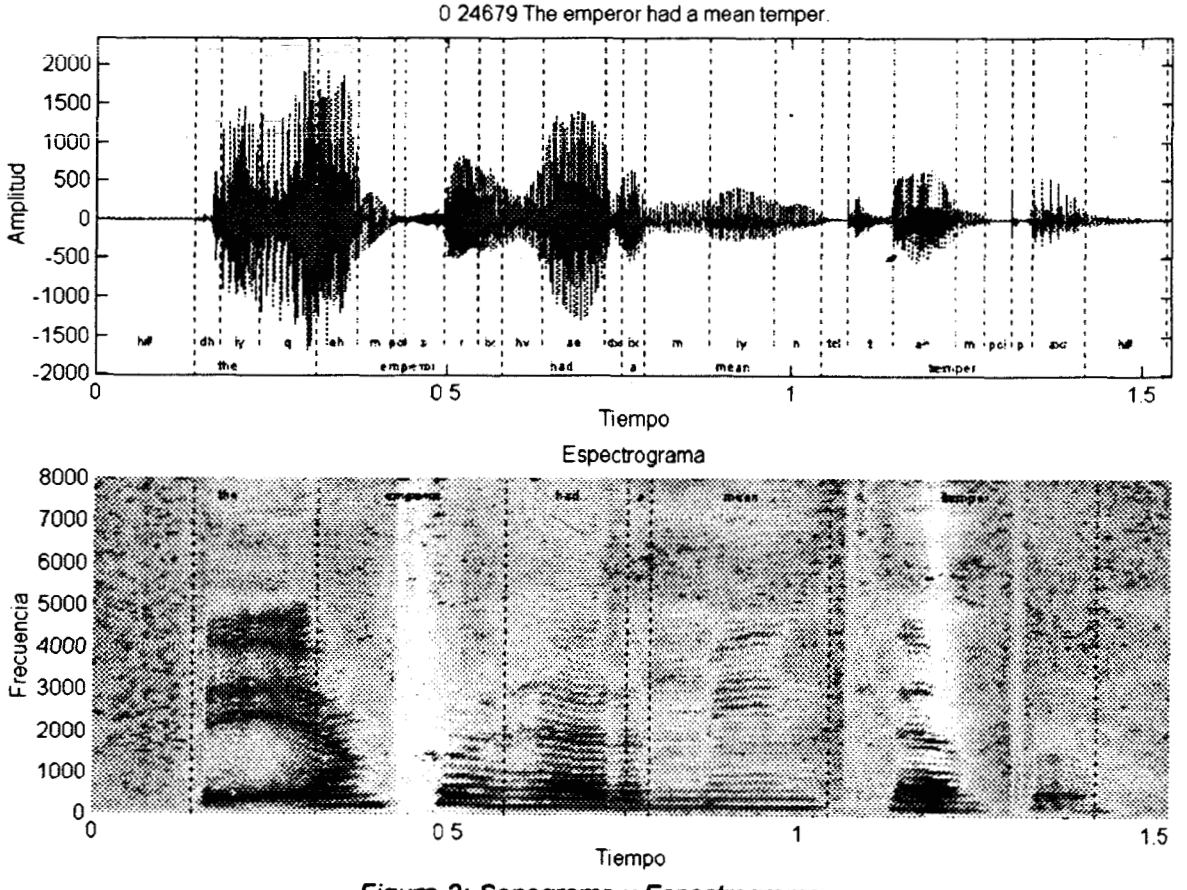

Figura 2: Sonograma y Espectrograma

No obstante lo anterior, actualmente se pueden desarrollar sistemas prácticos que utilicen solo algunas de las fuentes antedichas, ya que el sistema completo en el sentido anterior requeriría una cantidad enorme de procesamiento y sofisticación.

La simplificación más notable consiste en un sistema que reconozca solo palabras aisladas (generalmente 200 o 300 palabras) en un universo delimitado, o RPA y generalmente del tipo DH. En la otra punta tenemos el RDC sin restricciones en el hablante y aceptando frases complejas. Entre ambas hay una gran gama de configuraciones, algunas con aplicación industrial, sin embargo la ambiciosa meta del RDC parece resistirse a los más complejos métodos puestos en juego para abordarla.

El conocimiento fonético fue utilizado en los primeros sistemas de RPA. En estos sistemas se utilizaban plantillas fonéticas de referencia que se comparaban con la entrada al sistema, midiendo la distancia con respecto a ella se podía asociar dicha entrada con una palabra especifica. Como se puede ver, este enfoque tan rígido es imposible de extender al dominio del discurso continuo y de vocabulario ilimitado.

Dadas la alta variabilidad de las características acústicas de los sonidos según su contexto es necesario elegir una representación de la señal que incluya aspectos de la misma que puedan distinguir unidades elementales. Dos tipos de unidades en este sentido son: los alófonos y los fonemas. Los alófonos son las representaciones de los sonidos según aparecen realmente en las palabras. Los fonemas son representaciones más abstractas que capturan las características comunes de una clase de alófonos y se pueden caracterizar por una matriz de rasgos acústicos.

La mayor ventaja de la utilización de fonemas como unidad de base para representar las palabras habladas es que nunca hay más de 40 distintos por lengua (del orden de 20 para el castellano) ya que este tipo de representación no tendría en cuenta rasgos propios del hablante o emociones. La desventaja es que los fonemas son unidades abstractas que no se encuentran en forma explícita en la señal de voz. Es decir que, en el habla natural, los fonemas son realizados mediante la acción coordinada de todo el aparato fonador, por lo que existen fuertes modulaciones entre ellos. Por ello la señal acústica real de cada fonema depende de cual le precede, del que le sigue y del estado transitorio del aparato fonador.

Clasificar a los alófonos en fonemas abstractos necesita un análisis muy fino de la manera en la que el contexto del discurso determina los alófonos de un fonema. Estos están sometidos a alto grado de variación debido a las diferencias entre hablantes, las de un mismo locutor y las producidas por el contexto. La dependencia del entorno se debe al fenómeno ya mencionado de la coarticulación; es el caso por ejemplo de las características acústicas de las vocales que responden a las de las consonantes adyacentes y viceversa, y a si existen espacios a continuación o antes de ellas. Además, los rasgos de los alófonos pueden verse afectados por elementos suprasegmentales, es decir, que afectan a más de un segmento, como acento y entonación.

Otro tipo de unidad frecuentemente utilizada son los difonos, que consisten en la unión de dos fonemas desde la porción estable del primero hasta la posición estable del segundo. Esto asegura que se tiene en cuenta la información contenida en las transiciones, la cual ha demostrado ser de gran importancia para la inteligibilidad en pruebas psicoacústicas con voz sintética. En general han confirmado ser unidades fonéticas más fácilmente identificables que

los fonemas [GWS92]. Asimismo son mucho menos numerosos que las sílabas lo que facilita su manejo y la implementación de los métodos utilizados para su clasificación.

La dificultad para resolver los problemas asociados al reconocimiento del habla mediante técnicas de procesamiento convencionales, está dado por la complejidad de las señales implicadas, ya que las mismas presentan funciones estadísticas de densidad superpuestas, tienen formas complicadas en espacios de varias dimensiones o son no estacionarias [Koh88]. El castellano presenta ventajas comparativas frente a los idiomas anglosajones, y aún otras lenguas latinas, a los efectos de ser reconocido por una máquina. Su menor número de fonemas básicos, la mayor separabilidad de sus características acústicas y la correspondencia entre las regiones de decisión determinadas por estas últimas con los respectivos fonemas, facilitan la tarea; sin embargo, el problema más grave está centrado en aceptar las señales provenientes de cualquier hablante, con independencia del tono de voz, velocidad de pronunciación y discontinuidad en el modo de hablar.

Aunque las técnicas utilizadas por los sistemas de reconocimiento automático del habla han mejorado notablemente su desempeño en los últimos años, ésta dista mucho de ser adecuada para algunas aplicaciones. La mayor parte de los sistemas se basa en la aplicación de Modelos Ocultos de Markov (HMM), los que han sido útiles para tratar los aspectos secuenciales de la señal de habla, pero no han sido tan eficientes como clasificadores de fonemas [MoB95]. También se ha recurrido para esta tarea al empleo de redes neuronales artificiales y en particular a aquellas arquitecturas que permitan tratar los aspectos dinámicos de la señal de voz. La aparición de técnicas de entrenamiento eficaces para redes neuronales -en particular las redes anteroalimentadas- permitió la aplicación de las mismas al procesamiento del habla, aunque hasta hace poco tiempo estuvieron orientadas a patrones estacionarios. Para evitar este escollo se diseñaron redes neuronales que -además de los patrones estáticos- incorporaran simultáneamente información generada en diferentes instantes. Así surgieron las *Redes Recurrentes* (RNN's) y las *Redes con Retardos Temporales* (TDNN's) que permiten descubrir características acústico-fonéticas y sus relaciones a lo largo del tiempo [WHH89].

Descripción del Trabajo

Mediante el presente estudio se pretende avanzar un poco más hacia la compresión de los pasos implicados en el diseño de un dispositivo capaz de traducir voz a texto. El mismo consiste en una comparación objetiva entre distintos tipos de análisis o preprocesamiento para un sistema de RAH. Explícitamente la comparación se realiza entre el análisis de Fourier con ventanas y el análisis basado en Wavelets. Se supone que cuanto "mejor" sea el análisis o proceso utilizado para generar los patrones a identificar (en este caso de voz), más separadas quedan las clases en el espacio de patrones y las regiones obtenidas son más simples. Esto conlleva una mayor facilidad para aprender las regiones de decisión mediante técnicas de aprendizaje automático. En particular, distintas arquitecturas de redes neuronales artificiales han demostrado resolver bien los problemas relacionados con el RAH. Entre estas se pueden citar los *Perceptrones Multicapa* (MLP's) [RHW86], y como ya se mencionó las RNN's [Tak95], las TDNN's [WaH89], [WHH89]. así como también las *Redes Neuronales*

de Alto Orden (HONN's) [DaR95]. Estas pueden constituir una forma objetiva para medir el desempeño del análisis de acuerdo a la velocidad de aprendizaje y los errores cometidos en la clasificación. A pesar de ello es conveniente contar con algún método de evaluación más rápido y sencillo por lo que en este trabajo se explorarán también otras alternativas. Por otra parte las técnicas de aprendizaje no supervisado pueden brindar otra perspectiva acerca de la distribución de los patrones en el espacio generado por cada método [Koh88], [Lip87].

El tipo de análisis clásico para las señales de voz ha sido la *Transformada de Fourier de Tiempo Corto* (STFT) [RaS87], [Ope70]. También se pueden mencionar los métodos basados en *Coeficientes de Predicción Lineal* (LPC) [Mak75]. Sin embargo recientemente se ha desarrollado la *Transformada Wavelet* (WT) que permite realizar el análisis de señales no estacionarias en forma más "eficiente". Además se ha descubierto que el tipo de análisis realizado de esta forma es análogo al que realiza el oído a nivel de la cóclea [RiV91], [Dau92]. Esto nos alentaría a investigar esta herramienta debido a que nuestro oído es un dispositivo especialmente adaptado para el análisis de la voz. En realidad existe una adaptación recíproca entre el aparato fonador y el aparato auditivo y otras estructuras del Sistema Nervioso Central para asegurar la transmisión del mensaje contenido en la señal con la menor distorsión [Fle53]. Por otra parte existen diferentes bases o familias de Wavelets que pueden utilizarse para el análisis [RiV91], [AHT93], [Dau92] y habrá que resolver (de manera similar) cual es la base óptima para el caso planteado.

El idioma castellano (a diferencia del Inglés o Chino) ha demostrado ser más sencillo para su aprendizaje automático debido a que se pueden encontrar reglas únicas para su transcripción [Koh92], [Roc87], [GuB75]. Sin embargo el trabajo se realizará sobre una serie de fonemas del idioma inglés, en particular sobre las series más fácilmente confundibles. Esto se debe a que existen gran cantidad de bases de datos standard en este idioma, lo que permite conseguirlas fácilmente y poder comparar resultados con otras estrategias similares ya implementadas. Para los experimentos se elegirá el caso multi-hablante por ser el de aplicación práctica más directa. La razón de usar fonemas como unidad de clasificación se debe a su pequeña cantidad y a su gran difusión en el ámbito del RAH. Por otra parte parecen ser buenas unidades para la construcción de sistemas modulares para resolver el problema en forma más general (los difonos podrían ser otra opción) [WaH89].

Por último se debe mencionar que se han encarado comparaciones similares en otros ámbitos con resultados favorables para el Análisis Wavelets (por ejemplo [NaR95] para predicción del nivel de anestesia mediante Potenciales Evocados Auditivos). La diferencia entre los campos no permite extrapolar los resultados pero junto con otros factores ya mencionados alienta la experimentación con esta técnica. A pesar de que la señal de voz es una de las mejor estudiadas en el ámbito del procesamiento digital los trabajos basados en Wavelets se han orientado principalmente a cuestiones como compresión y filtrado pero prácticamente no ha sido empleada en sistemas de RAH [Fav94].

II . Aspectos Fisiológicos

Introducción

A los efectos de abordar esta comparación entre dos formas diferentes de análisis del habla sería conveniente conocer la señal de voz (objeto de nuestro análisis) y su forma de producción, de manera de comprender su naturaleza. Así mismo, deberíamos entender el procesamiento llevado a cabo por el sistema auditivo para discernir cuales son los parámetros relevantes de la señal que se deben extraer para lograr su reconocimiento. Todo esto vuelve imprescindible el estudio de los fundamentos anatómicos y fisiológicc involucrados en el proceso de emisión-percepción del habla.

Este capítulo se organizará de la siguiente forma. A continuación se describirá el mecanismo de producción del habla y los órganos involucrados. Esto incluye la descripción de los principales tipos de fonemas. Luego se presentarán aspectos relacionados con la señal de voz propiamente dicha mostrando algunos ejemplos. Finalmente se esbozarán los principios y elementos que intervienen en la audición.

Mecanismo de producción del habla

Para comenzar esbozaremos brevemente los mecanismos de involucrados en la producción del habla. Para un desarrollo detallado aplicado al idioma español remitirse a [Bor80], [RuZ92] o [Ruf94]. El aparato fonador se puede considerar como un sistema que transforma energía muscular en energía acústica. La teoría acústica de producción del habla describe este proceso como la respuesta de un sistema de filtros a una o más fuentes de sonidos. En la representación simbólica, si H es la función de transferencia del filtro que representa el tracto vocal en un instante dado y F la fuente de excitación, el producto $P = H \cdot F$ representa el sonido resultante. La fuente F indica la perturbación acústica de la corriente de aire proveniente de los pulmones. Se pueden identificar tres mecanismos generales en la excitación del tracto vocal:

1. Las cuerdas vocales modulan un flujo de aire que proviene de los pulmones dando como resultado la generación de pulsos cuasiperiódicos.

2. Al pasar el flujo de aire proveniente de los pulmones por una constricción en el tracto vocal se presenta la generación de ruido de banda ancha.

3. El flujo de aire produce una presión en un punto de oclusión total en el tracto vocal; la rápida liberación de esta presión, por la apertura de la constricción, causa una excitación de tipo plosivo, intrínsecamente transitoria.

En la Figura 3 se observa un esquema simplificado del aparato fonador y en la Figura 4 se aprecia una sección sagital del mismo. La zona comprendida entre la laringe (glotis) y los labios constituye el tracto vocal propiamente dicho.

En síntesis, los sonidos del habla son el resultado de la excitación acústica del tracto vocal por la acción de una o más fuentes. En este proceso los órganos fonatorios desarrollan distintos tipos de actividades, tales como movimiento de pistón que inician una corriente de aire, movimiento o posiciones de válvula que regulan el flujo de aire, y al hacerlo generan sonidos o en algunos casos simplemente modulan las ondas generadas por otros movimientos.

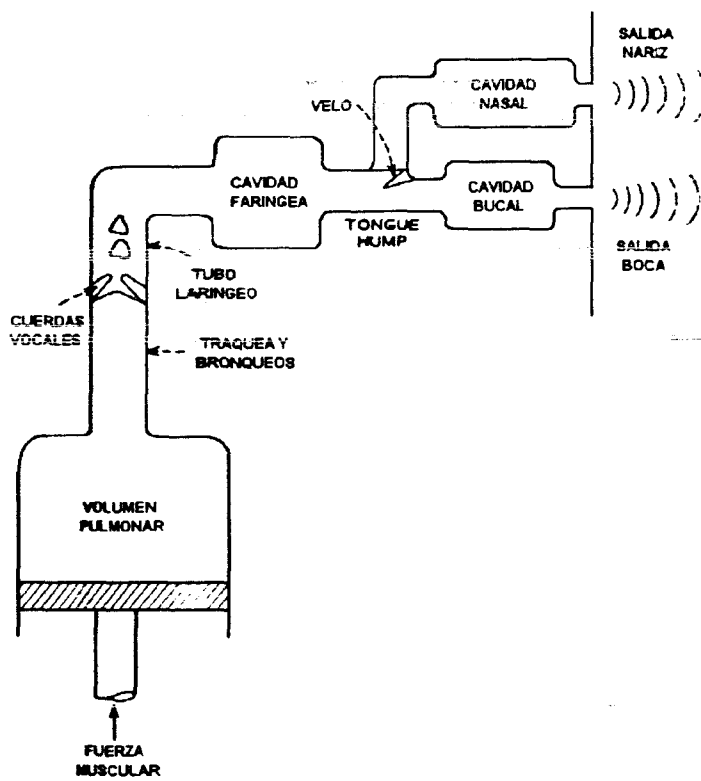

Figura 3: Esquema del aparato fonador

El sistema respiratorio constituye la principal fuente de energía para producir sonidos en el aparato fonador humano. La energía es proporcionada en forma de flujo o corriente de aire y presiones que, a partir de las distintas perturbaciones, generan los diferentes sonidos. El aparato respiratorio actúa también en la regulación de parámetros tan importantes como la energía (intensidad), la frecuencia fundamental de la fuente periódica, el énfasis y la división del habla en varias unidades (sílabas, palabras, frases).

La laringe juega un papel fundamental en el proceso de producción del habla. La función fonatoria de la laringe se realiza mediante un complejo mecanismo en el que intervienen no sólo los pliegues vocales, los cartílagos en los que se insertan y los músculos laríngeos intrínsecos sino también de las características del flujo de aire proveniente de los pulmones. La forma de onda de los pulsos generados puede representarse como una onda triangular.

El tracto vocal está formado por las cavidades supraglóticas, faríngeas, oral y nasal, como se ilustra en la Figura 5.

El tracto vocal puede mantener una configuración relativamente abierta y actuar sólo como modulador del tono glótico o estrechar o cerrar el paso de la corriente de aire en una zona específica. El tracto actúa como filtro acústico, principalmente en los sonidos con componente glótica, pudiendo modificar sus parámetros en forma continua. Si se observan los espectros de los sonidos vocálicos, éstos proporcionan información sobre todos los

aspectos relevantes de la configuración del tracto en ese instante. Es decir, todas las resonancias del tracto, resultantes de su configuración, pueden observarse directamente en el espectro del sonido vocálico.

Consideramos ahora las configuraciones del tracto que corresponden a cada sonido ya que -como se dijo antes- toda configuración presenta características propias de resonancia que junto con la fuente de excitación actuante, dan al sonido su peculiar cualidad fonética. Por ello se agrupan los sonidos en vocálicos y consonánticos. Esta división se sustenta tanto en las características acústicas como en los gestos articulatorios que dan lugar a cada tipo de sonido.

En la articulación de vocales y sonidos tipo vocálicos, el tracto presenta una configuración relativamente abierta y la fuente de excitación es siempre glótica. Las propiedades de estos sonidos persisten por un tiempo apreciable o cambian muy lentamente mientras se mantenga la configuración del tracto.

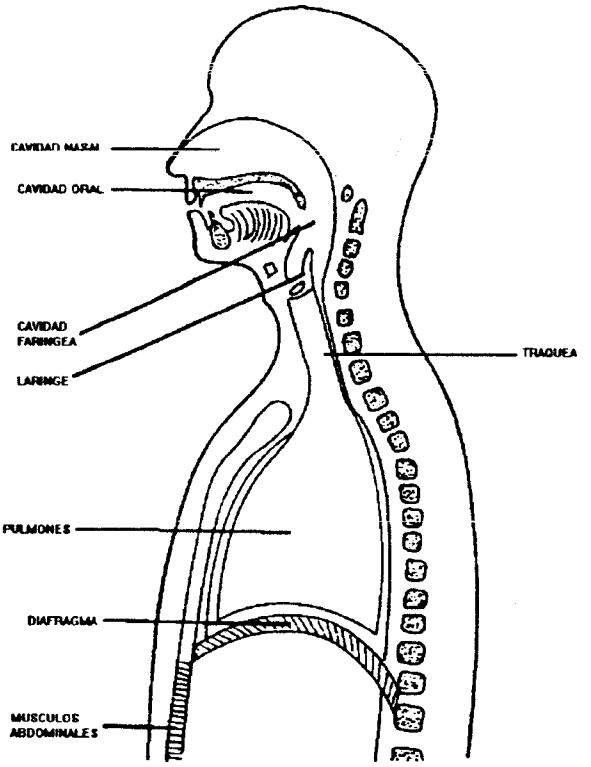

Figura 4: Corte del Aparato Fonador

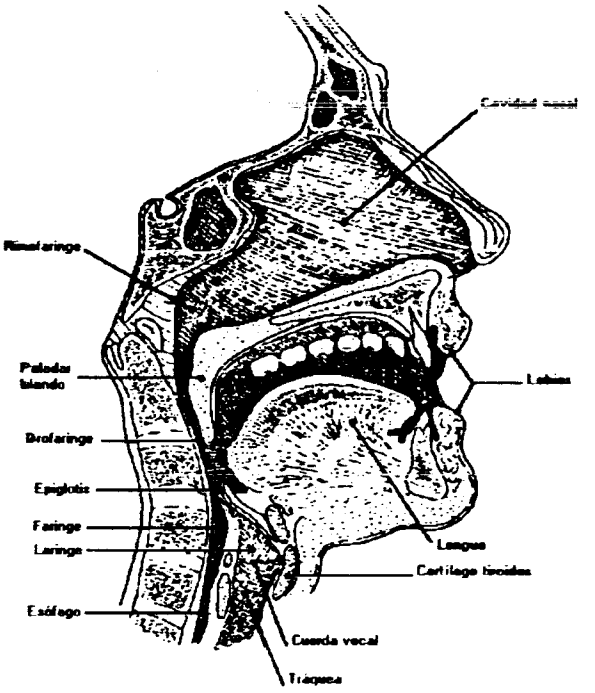

Figura 5: Cavidades Supraglóticas

Los pulsos glóticos estimulan el tracto vocal que actúa como sistema resonador. Este puede variar su configuración y con ello sus frecuencias de resonancias. Esta posibilidad de variación es la que permite al hablante producir muchos sonidos diferentes. La forma del tracto en la producción de las vocales esta controlada principalmente por la posición de la lengua, de la mandíbula y de los labios. Se pueden clasificar los sonidos vocálicos por distintas características acústicas:

- Zonas de estrechamiento: Por estudios sistemáticos de radiografías de articulaciones vocálicas se han localizado cuatro zonas de localización de la constricción, de esta manera los sonidos vocálicos se agrupan en Palatales (/i/, /e/), Velares (/u/), Velofaríngeos (/o/) y Faríngeos (/a/) según el lugar de la constricción.

- Grado de estrechamiento: De esta manera se describen los sonidos vocálicos según el grado de estrechamiento en la región de menor área o constricción máxima, Constricción estrecha (/i/, /u/, /o/), amplia (/e/, /a/).

- Abertura de la boca: Esta abertura cuya configuración y grado están determinadas por la acción de los labios y del maxilar inferior, da lugar a importantes diferenciaciones acústicas y fonéticas. Abertura amplia (/a/), abertura más reducida (/i/, /u/).

- Longitud del tracto: La longitud del tracto se modifica redondeando los labios, subiendo y bajando la posición de la laringe. Labializado (/o/, /u/), deliabializado (/a/).

Los sonidos consonánticos se producen con una configuración relativamente cerrada del tracto vocal. El cierre o estrechamiento del canal se realiza en zonas especificas del tracto vocal por acción de partes especificas de las estructuras articulatorias. Entre los factores que determinan la cualidad del sonido resultante, debemos distinguir aquellos que hacen al modo de articulación (cierre o estrechamiento) de los que señalan la zona o lugar de articulación (lugar donde se produce cierre o estrechamiento). La participación de la fuente glótica, la naturaleza del cierre o estrechamiento y la transmisión a través de la cavidad oral y/o nasal, constituyen los principales factores del modo de articulación.

Las consonantes, por otro lado, pueden ser agrupadas en los siguientes tipos articulatorios:

- Oclusivas : se producen por el cierre momentáneo total o parcial del tracto vocal seguido de una liberación más o menos abrupta del aire retenido. Por ejemplo las totales /p/, /t/, /k/ o las parciales /b/, /d/, /g/.

- Laterales : estos se producen cuando se hace pasar la señal sonora glótica por los costados de la lengua. Por ejemplo /l/ y /ll/.

- Nasales : son producidas a partir de excitación glótica combinada con la constricción del tracto vocal en algún punto del mismo. Por ejemplo /m/, /n/.

- Vibrantes : estos son producidos al pasar el aire por la punta de la lengua y producir su vibración. Tienen componente glótica. Por ejemplo /r/ y /rr/.

- Fricativas : se caracterizan por ser ruidos aleatorios generados por la turbulencia que produce el flujo de aire al pasar por un estrechamiento del tracto. Pueden ser sonoros como /y/ si hay componente glótica o sordos como /f/, /s/ o /j/ si no la hay.

- Africadas : Si los fonemas comienzan como oclusivos y la liberación del aire es fricativa se denominan africados. Por ejemplo la /ch/ del castellano.

- Semivocales : están formadas por la unión de dos de los anteriores hasta el punto de convertirse en otro sonido. Algunos consideran en este grupo a las vibrantes y las laterales, así como también la /w/ del inglés.

De lo dicho anteriormente, se podría inferir que el habla es, de alguna manera, un hecho discreto, es decir una sucesión de sonidos vocálicos y consonánticos. Pero si observamos la señal de la voz, la representación acústica de una frase, veremos muy pocas pausas o intervalos entre los sonidos. El habla constituye un continuo acústico, producido por un movimiento ininterrumpido de algunos órganos del aparato fonador. A pesar de la naturaleza continua de la voz los oyentes pueden segmentarlas en sonidos.

Las características suprasegmentales de la voz están determinadas por la entonación, la cual determina la prosodia. Las variables que intervienen en la entonación son las variaciones de frecuencia fundamental o pitch, la duración y variaciones de energía sonoridad.

La prosodia en las uniones puede ser caracterizada por silencios, duración en las vocales, o por formas como puede ser la presencia de sonoridad o aspiración. Por ejemplo en la frase "perdonar, no matar" existe una pausa después de "perdonar" pero si la coma cambia de lugar "perdonar no, matar" el silencio se produce después de "no" cambiando totalmente el significado del mensaje.

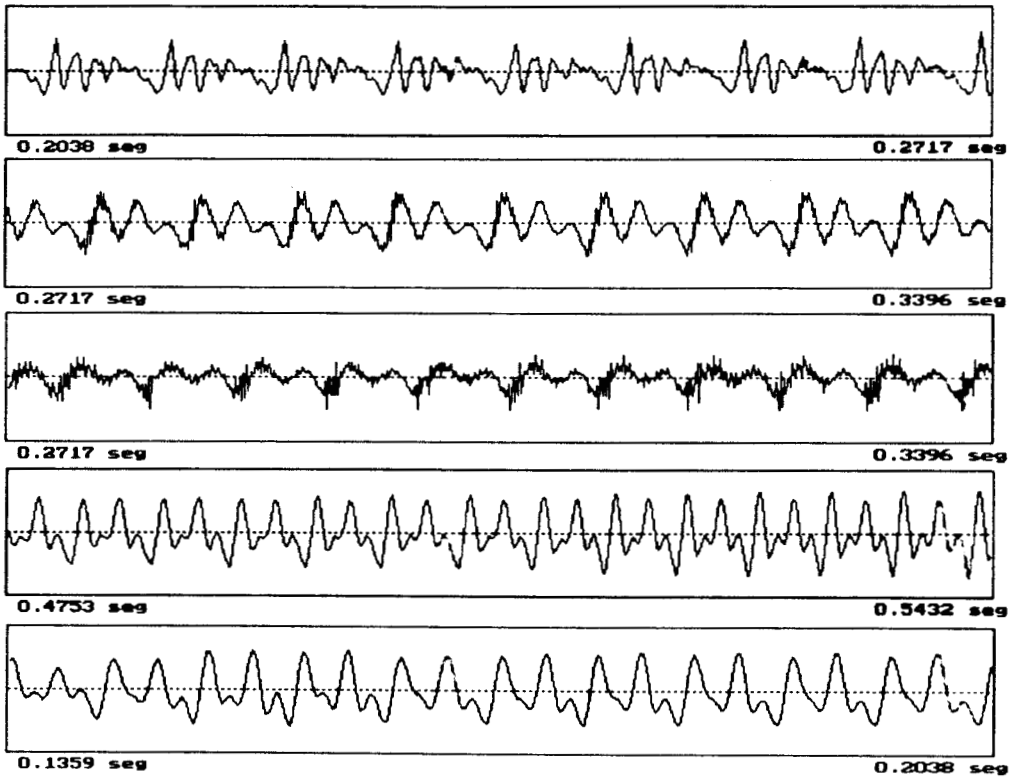

Figura 6: Sonograma de /a/, /e/, /i/, /o/, /u/ (español)

La Señal de Voz

Hasta ahora hemos descripto los distintos tipos de fonemas y la forma en la que se originan en el aparato fonador. Sin embargo hemos hecho pocas referencias a los aspectos referentes a la señal o su espectro que son de alguna manera el substrato del que obtendremos nuestros patrones para pasar al clasificador.

Empezaremos por analizar las vocales. en la Figura 6 se observa el sonograma de las vocales del español pronunciadas en forma sostenida y aislada por un hablante femenino. En este caso se observa un fuerte parecido entre /o/ y /u/, lo cual es de suponer porque se puede decir que son vocales 'cercanas'.

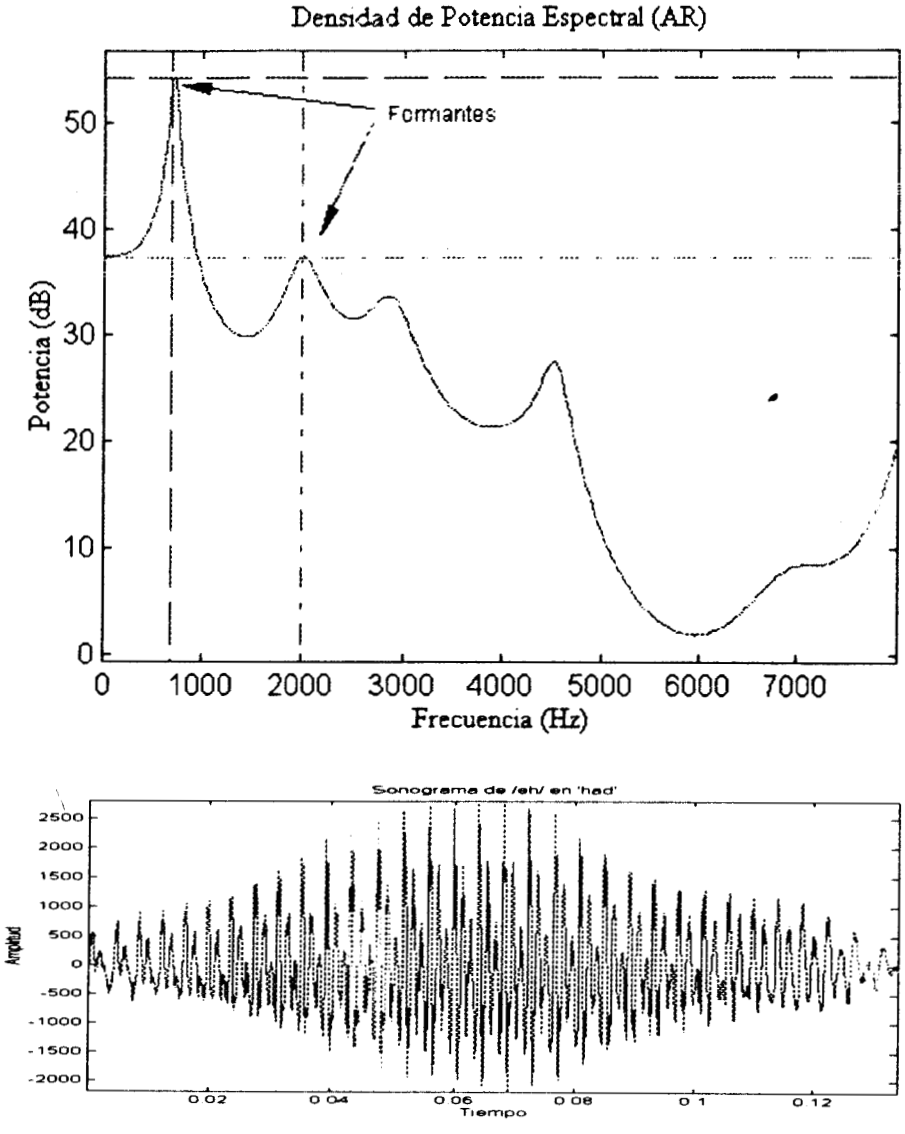

Figura 7: Sonograma y Espectro de una /eh/

Como ya se dijo en los espectros de los sonidos vocálicos pueden observarse todas las resonancias del tracto. Estas resonancias aparecen como picos en el espectro y se denominan formantes. En la Figura 7 aparece el sonograma de una /eh/ en la palabra inglesa 'had' y el espectro suavizado (estimado con un modelo AR) donde se aprecian claramente los picos. Las formantes se numeran a partir del 0, correspondiendo f0 a la frecuencia fundamental directamente relacionada con la entonación de una frase o emisión. El resto de las formantes, principalmente f1 y f2, constituyen un medio para caracterizar a las vocales. En la Figura 18 (Capítulo siguiente) se pude apreciar un gráfico de distribución de las vocales inglesas en función de f1 y f2. La presencia de formantes evidencia si se trata de un trozo sonoro o sordo (con o sin componente glótica).

Existen algunas características de la señal de voz que se pueden evidenciar mediante análisis relativamente sencillos como ser la *Energía de Corto Tiempo* y la *Cantidad de Cruces por Cero*. Estos análisis tienen la ventaja de ser sencillos en su implementación digital y muy rápidos. La Energía da una idea de la intensidad de la señal en función del tiempo y constituye una parámetro de suma importancia ya que permite diferenciar entre varios tipos de fonemas y constituye una parte esencial de la entonación (junto con f0). Los Cruces por Cero constituyen una medida indirecta del contenido frecuencial de la señal.

En la Figura 8 se observan estas curvas para la palabra inglesa "suit". La curva de Cruces

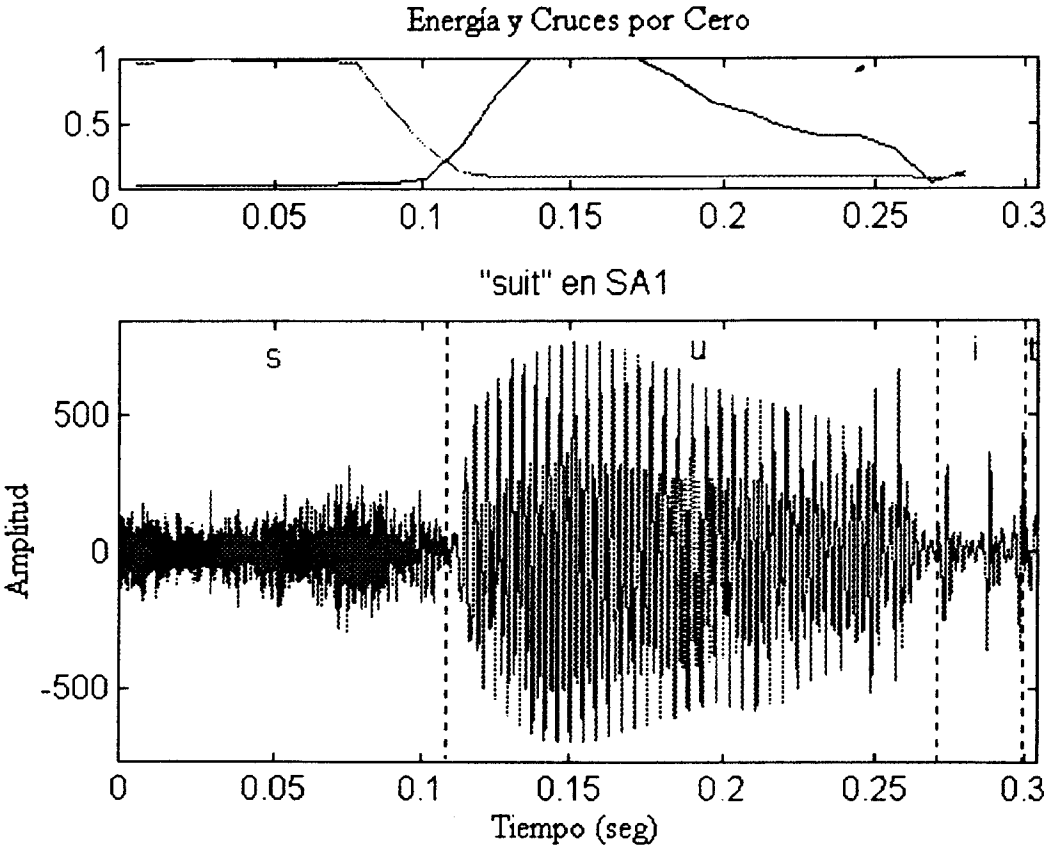

Figura 8: Sonograma, Cruces por Cero y Energía de 'suit'

por Cero se muestra en color violeta y la de Energía en color Azul. La primera porción de la palabra corresponde al fonema fricativo /s/, lo que se refleja en el sonograma como un trozo ruidoso. Así mismo la cantidad de Cruces por Cero es muy alta debido al alto contenido frecuencial de la señal. Por el contrario la energía es relativamente baja en relación al trozo correspondiente a la vocal. De esta manera ambos análisis permiten distinguir rápidamente entre fonemas sonoros y sordos.

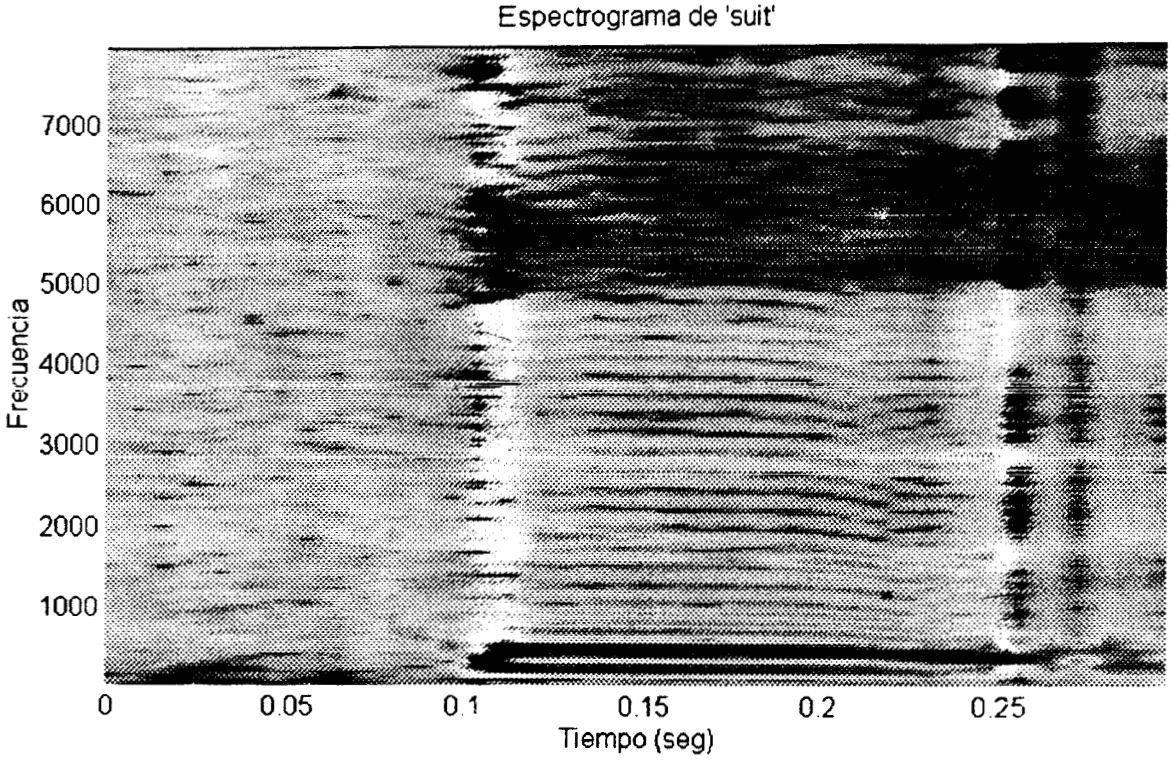

Figura 9: Espectrograma de la palabra 'suit'

En la Figura 9 se observa el espectrograma de la misma emisión donde se aprecia también el contenido de alta frecuencia de la /s/, la estructura formántica de /u/,/i/ y la corta duración de la oclusiva /t/.

Podrían llenarse muchas páginas con gráficos y análisis de los distintos fonemas. Sin embargo nuestro interés aquí no es presentar este material de manera exhaustiva sino más bien, y como ya se mencionó, mostrar unos pocos ejemplos que permitan comprender mejor la naturaleza de la señal de voz.

Fisiología de la Audición

En nuestro trabajo, el interés principal es comprender como se realiza el procesamiento de la señal de habla por la periferia auditiva, en particular el procesamiento del castellano hablado y otros lenguajes fonéticamente similares.

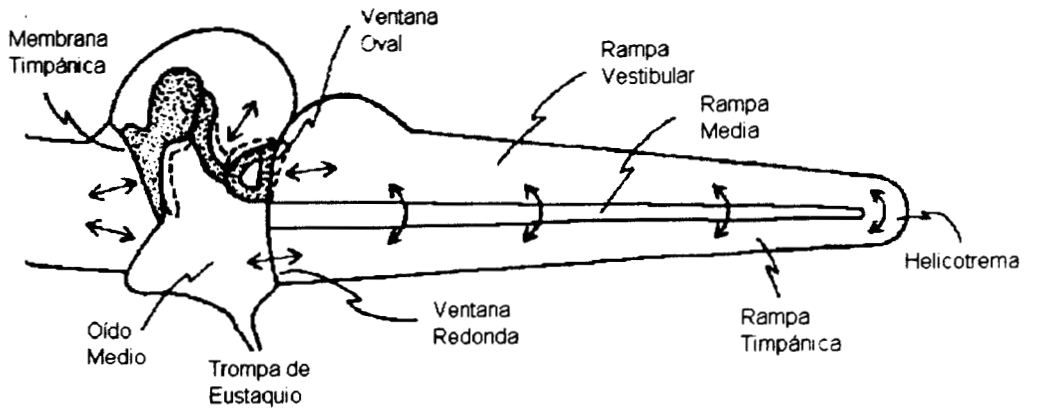

Figura 10 : Corte del Oído

Nos podríamos hacer la siguiente pregunta: ¿Que propiedades de la periferia auditiva son particularmente apropiadas para codificar la voz?. La respuesta en parte, se encuentra en la capacidad magnífica del sistema auditivo para resolver simultáneamente tanto las características espectrales como temporales de estímulos de banda ancha.

En la Figura 10 puede apreciarse un corte transversal del oído. En el se observa parte del oído externo, el oído medio y el interno (la cóclea se halla desplegada para mayor claridad). En la Figura 11 puede apreciarse un corte de la coclea y en la se aprecia una ampliación del mismo donde se distinguen las células ciliadas encargadas de la transducción mecánico-eléctrica y la membranas Basilar y Tectoria. Para una descripción anatómica y fisiológica detallada de la vía auditiva remitirse a la extensa bibliografía al respecto, como por ejemplo [Som86].

El oído humano funciona en un medio aéreo y por ello es comprensible que represente un aparato bastante eficiente para la recepción de sonidos transmitidos por el aire.

Cuando funciona normalmente, es estimulado por las ondas de presión que se transmiten a través del aire siguiendo el conducto auditivo externo hasta el tímpano o membrana timpánica, cuya superficie es de aproximadamente 70 mm^2.

En el funcionamiento normal del sistema auditivo el sonido se transmite desde la membrana del tímpano a través de la cadena de huesecillos del oído medio, cuya función principal es adaptar impedancias [KhT72]. El más interno de ellos -el estribo- establece contacto con la ventana oval que esta ubicada en la base de la cóclea.

Una vez excitada la ventana oval el sonido se transmite a través de la perilinfa de la rampa vestibular en la cóclea, atraviesa el helicotrema y sigue su recorrido en la rampa timpánica hasta la ventana redonda.

La ventana oval y la redonda trabajan de forma tal que cuando una se comba hacia adentro la otra se comba hacia afuera y viceversa, el movimiento hacia adentro y afuera se repite con la misma frecuencia del estímulo sonoro.

En la cóclea es donde tiene lugar la transducción. Ésta se produce como respuesta a una curvatura de las cilias de las células ciliadas, esta curvatura produce una variación en el potencial de membrana de las células; si las cilias se curvan hacia el cuerpo basal se produce una despolarización, mientras que si se curvan en el otro sentido se produce una hiperpolarización.

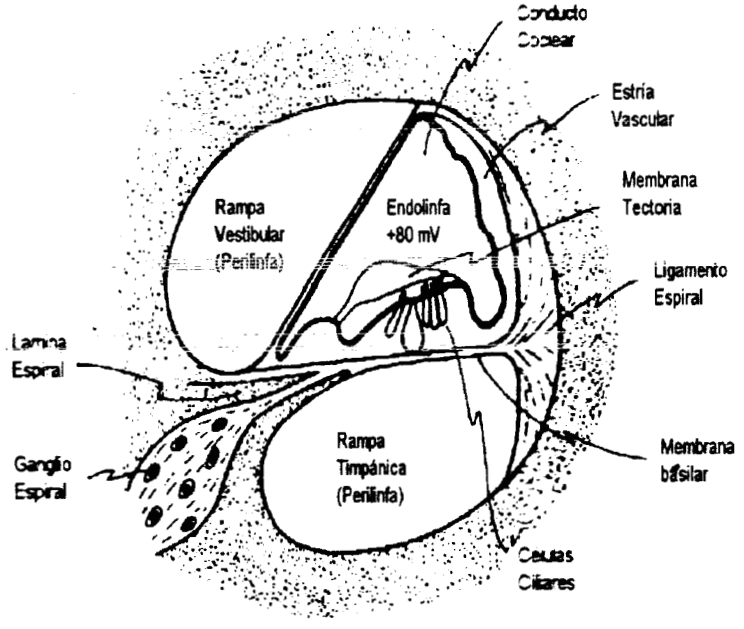

Figura 11 : Corte de la Cóclea

La excitación de las células ciliadas está determinada, en gran medida, por las excursiones de la membrana Basilar; sobre la cual actúan las ondas de presión oscilatorias resultantes de la transmisión del sonido en las rampas vestibular y timpánica.

Hasta el trabajo de Georg von Békesy en los 1940s y 50s [Bék60], los conceptos de vibración de la membrana Basilar se basaron en anatomía [Hel54] o estudios psicofísicos [Fle53]. Debido a los niveles altos de intensidad utilizados y a que la mayoría de sus experimentos utilizaban huesos temporales de cadáver, sus resultados nos dan solo una imagen de primer orden de la mecánica coclear.

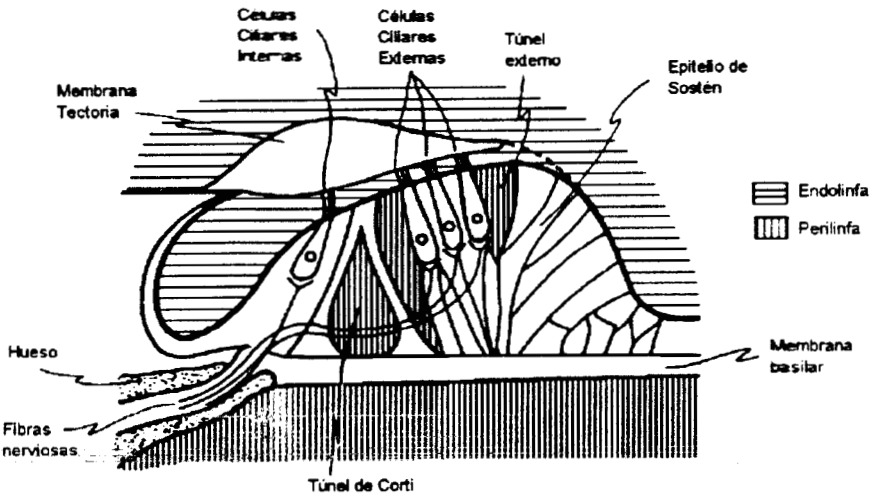

Figura 12: Detalle de las Células Ciliadas.

El hecho de que la membrana Basilar sea más rígida en un extremo que en el otro tiene una consecuencia muy importante; la membrana es más rígida cerca de la ventana oval donde su ancho es mínimo, por lo tanto tiene menor cantidad de masa por unidad de longitud; estas características hacen que la membrana en esta región vibre con preferencia ante un estímulo de alta frecuencia. De esta forma, las vibraciones de alta frecuencia tendrán su máxima amplitud cerca del lugar donde las ondas comienzan a desplazarse, pronto disiparan la mayor parte de su energía y se desvanecerán en el camino no alcanzando nunca el vértice. Las vibraciones de baja frecuencia, por el contrario, comenzaran con una amplitud pequeña cerca de la base y la irán aumentando a medida que se acerquen al vértice; de esta forma tenemos representadas las frecuencias audibles a lo largo de toda la cóclea. De esta manera, la amplitud de las

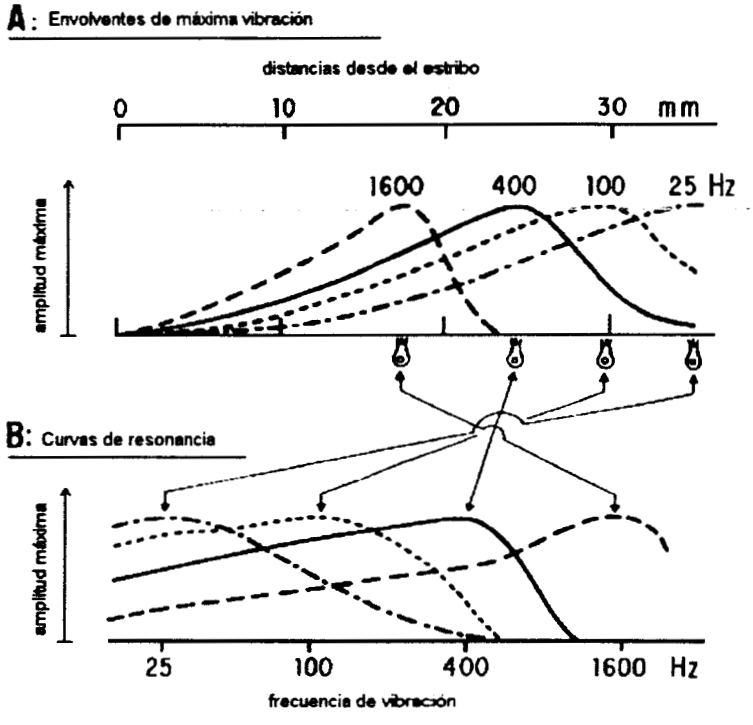

Figura 13: Envolventes de Máxima Vibración y Curvas de Resonancia.

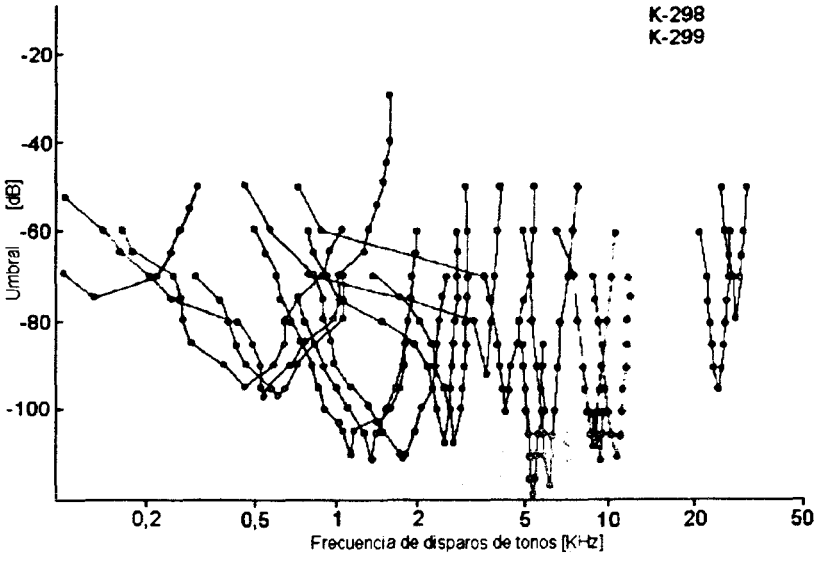

Figura 14 : Curvas de Sintonía Nerviosa

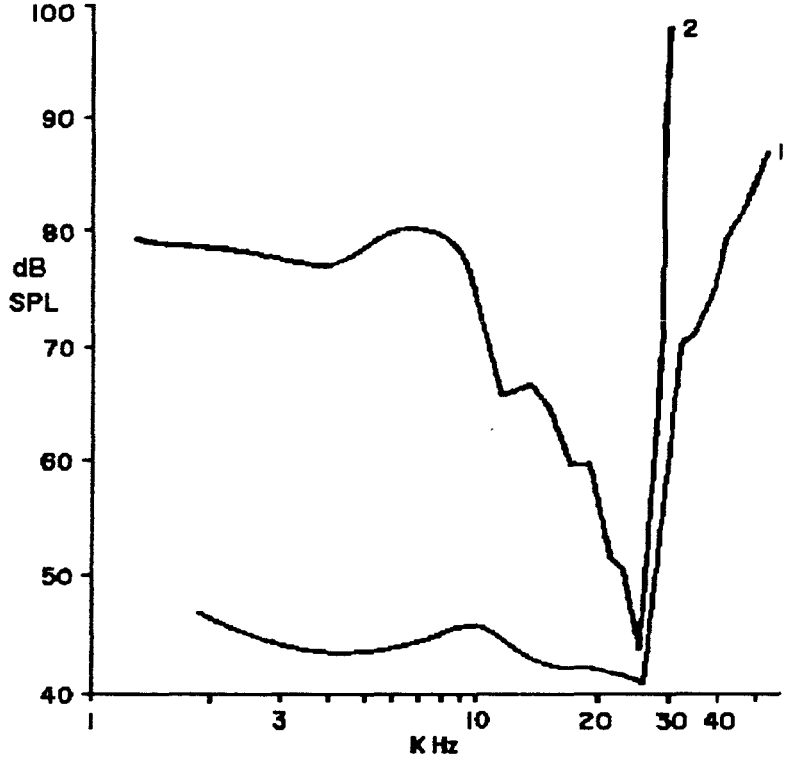

Figura 15 : Resonancia Mecánica y Sintonía Nerviosa

vibraciones en distintos puntos de la cóclea varía con la frecuencia del estímulo; el grado en el cual es excitada una determinada célula ciliada es una función conjunta de su posición en la membrana Basilar y de la amplitud del estímulo.

Las excursiones máximas de la membrana Basilar han sido mapeadas como una función de la distancia al estribo, para tonos de igual intensidad pero distintas frecuencias. Estos mapas se denominan envolventes de la onda de desplazamiento (**¡Error! No se encuentra la fuente de la referencia.**-A). Empleando los datos necesarios para la construcción de estas envolventes, también se pueden graficar las amplitudes relativas de las excursiones para los distintos puntos sobre la membrana Basilar como una función de la frecuencia del estímulo; estas son las curvas de sintonía mecánica o curvas de resonancia (**¡Error! No se encuentra la fuente de la referencia.**-B).

La curva de resonancia de la membrana Basilar describiría con precisión la excitación de las células ciliadas en función de la frecuencia, si éste fuera el único factor que influyera

en la vibración de las células ciliadas. Sin embargo, las propiedades mecánicas de las cilias y de la membrana Tectoria que las cubre también influyen en la vibración de las células ciliadas; de hecho, la rigidez de las cilias. la masa y la elasticidad de la membrana Tectoria varían de un extremo al otro de la cóclea.

Estas características del complejo célula-membrana Tectoria tiene el efecto de limitar la sintonía de las células ciliadas a un ancho de banda de frecuencias más estrecho que el del punto de la membrana Basilar donde se encuentra la célula.

En los 1960s, los estudios fisiológicos de trenes de impulsos en fibras de nervio auditivo únicas de gatos [KWT65] proveyeron nueva información y se suscitaron mas preguntas acerca de la periferia auditiva. Al comienzo de estos estudios se hizo uso extensivo de "tonos puros" (sinusoides).Una vez que se "aisló" una fibra se pudieron registrar impulsos de esa fibra única. Era usual obtener una "curva de sintonía" que trazaba las respuestas umbral versus la frecuencia (Figura 14). El mínimo de una curva de sintonía indica el lugar a lo largo del caracol que ocupa la célula ciliada que excita la fibra. Una característica de las curvas de sintonía es que cerca de su mínimo (frecuencia característica o C.F.), su forma es mucho aguda que las curvas de sintonía de los resultados de Bekesy.

En la Figura 15 vemos la curva de resonancia en un punto de la membrana Basilar (1) y la curva de sintonía de una fibra nerviosa que inerva a la célula ciliada en ese punto (2); la curva de resonancia (1) muestra los niveles de presión sonora relativos requeridos para hacer vibrar la membrana en ese punto a una amplitud dada para varias frecuencias de sonido, la curva de sintonía (2) muestra el umbral de la fibra nerviosa a los estímulos sonoros de frecuencia variable. Nótese que las curvas (1) y (2) tienen frecuencias de corte similares, pero del lado de las bajas frecuencias la curva (2) es mucho más escarpada que la curva (1). Varios mecanismos fueron propuestos para explicar esta aparente discrepancia entre las curvas de sintonía mecánicas y neurales. Estudios de la mecánica de la membrana Basilar utilizando métodos refinados mostraron una agudeza de sintonía mecánica bastante parecida a la de la sintonía neural [Rug92].

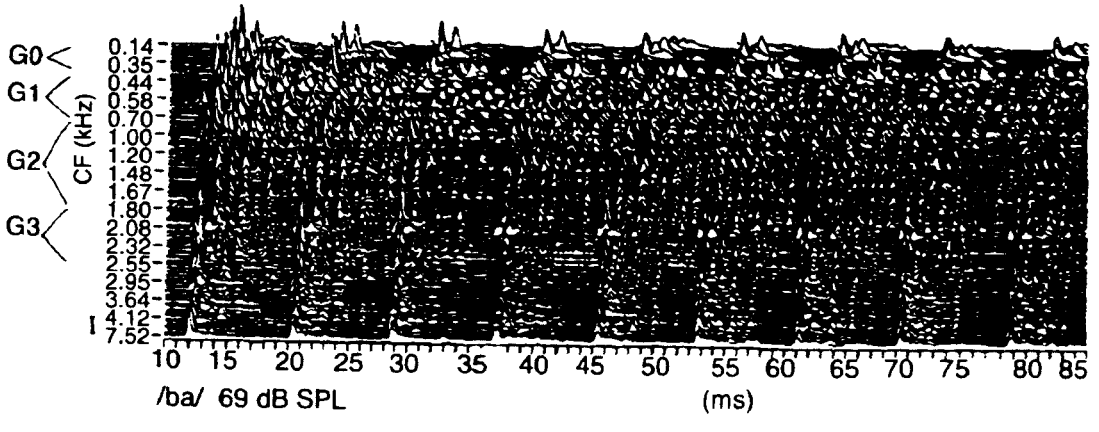

Figura 16 : Neurograma de /ba/

Se puede decir que la membrana Basilar esta mecánicamente sintonizada con la frecuencia del sonido aplicado, por esta razón se puede pensar que las descargas nerviosas provenientes de zonas determinadas de la membrana Basilar ya tienen la información de la frecuencia del estimulo. El estudio detallado de la respuesta del nervio auditivo a estímulos complejos tales como el habla esta recién en sus comienzos. Aunque no hay un camino para dar una descripción definitiva en este momento, es razonable asumir que la codificación de la señal del habla por el nervio auditivo esta caracterizada por contener un numero finito de elementos (aproximadamente 30.000 fibras del nervio auditivo en el hombre) y las respuestas de cada elemento están determinadas por una secuencia compleja de estados distribuidos e iterativos que preceden la iniciación de las espigas de descarga. Las fibras en la cóclea están tonotópicamente organizadas, de modo que las neuronas que inervan la base responden mejor a las altas frecuencias y aquellas que están en el ápice a las bajas frecuencias. Así el nervio auditivo puede considerarse una disposición ordenada de elementos arreglados de acuerdo a la frecuencia característica. Los elementos -fibras- en esta disposición responderán incrementando su probabilidad de descarga cuando el nivel del estimulo supera el umbral.

Para sonidos vocálicos intensos es posible que todas las fibras con frecuencias características debajo de 4-5 KHz puedan estar en estado de saturación. Bajo estas circunstancias, la información acerca de la frecuencias formantes a niveles del habla ordinario pueden no aparecer en la distribución espacial de información de frecuencia promedio y deben ser codificadas en forma diferente, probablemente en patrones temporales de descarga. En particular, los patrones de respuesta de fibras con frecuencia característica cercana a la frecuencia formante pueden ser dominadas por componentes sincronizadas de esa frecuencia formante.

El neurograma [SWS90] es una representación directa de la información experimental o la salida de un modelo del oído. Un neurograma basado en respuestas fisiológicas al sonido CV (consonante-vocal) sintetizado /ba/ se muestra en la Figura 16. Cada línea del neurograma representa tasa de disparo instantánea promedio (IFR) de una fibra nerviosa. La C.F. de la fibra está dada a la izquierda. A pesar del parecido con el clásico espectrograma el neurograma presenta información de manera distinta, utilizando otra forma de codificar los patrones generados "más a la medida del sistema auditivo".

Otro dato neurobiológicos importante es la arquitectura neuronal de la corteza auditiva. La corteza está formada por varias capas de células nerviosas, cada una de las cuales está constituida por tipos específicos de neuronas. Las capas corticales superiores fueron desarrolladas en las últimas etapas evolutivas del cerebro de los mamíferos y, en el caso del hombre, poseen una gran proporción de la totalidad de las neuronas [Sej86], [Kel85], [Mar85], [LaZ87]. La actividad neuronal sigue, en general, un patrón vertical que da lugar a la formación de columnas que a su vez están relacionadas lateralmente entre sí. Dentro de cada columna, una neurona perteneciente a una capa hace sinapsis directas sobre neuronas de la siguiente capa, o bien indirectamente, a través de interneuronas [Mar85]. Esto da lugar -teniendo en cuenta los retardos sinápticos- a que una neurona cualquiera de las capas más altas reciba simultáneamente información que fue generada en instantes distintos en la periferia, lo que permite establecer relaciones temporales complejas.

III . Los Datos

Introducción

Gran parte del presente trabajo depende de los datos o muestras de voz etiquetadas (corpus) utilizados para generar los patrones para entrenamiento y prueba de los clasificadores. La gran influencia sobre los resultados del corpus empleado tiene principalmente tres razones:

1. La cantidad de emisiones, cantidad de hablantes y diversidad fonética determinan la complejidad de la tarea de reconocimiento.

2. De la fiabilidad de los mismos depende en gran medida la validez de los resultados obtenidos.

3. La disponibilidad y difusión de la base de datos empleada repercute sobre la posibilidad de comparación con otras estrategias pasadas o futuras.

De acuerdo con estos puntos se decidió utilizar una base de datos standard del tipo citado en la bibliografía especializada y los artículos más recientes del área. Este enfoque, como se mencionó, tiene la ventaja de poder comparar los resultados con los reportados previamente. Sin embargo estas bases de datos están disponibles generalmente en idioma inglés y los resultados no son directamente extrapolables al español (en general se esperaría mejor desempeño que para el inglés). Estos corpora se pueden obtener en CD-ROM a través de distintas instituciones (por ejemplo [LDC], [OGI]). Existen dos bases de datos muy utilizadas, una de ellas es TIMIT [FWD86], [MoB95] para discurso continuo y la otra es NIST TI-46 [FaK94], [Fav94] para palabras aisladas. Esta última es una base multi-hablante y se han reportado gran cantidad de resultados para un subconjunto denominado E-set, por tratarse de un conjunto de palabras altamente confundible entre sí. Este conjunto está compuesto por las palabras correspondientes al alfabeto inglés que tienen como segundo fonema a /e/ (como /be/, /de/, /ge/, etc.). La baja energía y corta duración de las consonantes en relación a la vocal hacen que sea un conjunto difícil de clasificar. La base TIMIT es también multi-hablante, pero bastante más grande en tamaño, y es una de las más empleadas en el ámbito del discurso continuo por ser la más grande, completa y mejor documentada de su tipo. Ya que nuestro trabajo está orientado principalmente al reconocimiento de fonemas en discurso continuo se eligió TIMIT. Esta base o corpus posee una gran cantidad de fonemas en diversos ambientes y pronunciados por más de 600 hablantes diferentes. Esto constituye un total de unas 5 horas de material hablado etiquetado y casi 650 MBytes de información para procesar, lo que da una idea de las dificultades involucradas en su manipulación y utilización, así como también de la complejidad de la tarea de clasificación de los fonemas contenidos en la señal. Las dimensiones del problema del RAH se plantearon en el capítulo de introducción y se podría decir desde este punto de vista que se trata de un problema de identificación de fonemas en discurso continuo e independiente del hablante. Otro punto importante de mencionar es que, debido a los procedimientos de grabación, la

señal registrada está prácticamente libre de ruido, situación que dificilmente se de en condiciones distintas a las de laboratorio.

A pesar de que se dispone de la infraestructura necesaria para el diseño e implementación de una base similar a TIMIT en idioma castellano esto puede requerir de gran cantidad de recursos humanos y tiempo, lo que hace imposible su uso en el presente proyecto. Para futuros trabajos se pretende confeccionar esta base. Para ello se utilizará una cámara anecoica diseñada para este tipo de registros disponible en el Laboratorio de Audiología de la UAMI. Las emisiones serán procesadas y etiquetadas por un software que se está desarrollando al efecto [ARZ93], [ARZ94], [Aru94]. El diseño del corpus se realizará en forma conjunta con el departamento de Lingüística de la UAMI, con el que se ha comenzado ha colaborar en el marco de un proyecto interdisciplinario.

En lo que se sigue se describirán las características principales de TIMIT, así como su organización y tipos de archivos, ilustrándose con algunos ejemplos. Luego se explicará como se escogieron los hablantes y se detallarán las condiciones utilizadas durante la grabación de los datos. Siguiendo se describirá el tipo de texto empleado en las oraciones y la separación de los datos en entrenamiento y prueba. A continuación se presentarán los fonemas registrados y los símbolos utilizados para representarlos. Por último se explicarán los criterios empleados para elegir el conjunto de emisiones que emplearemos en el trabajo.

Descripción de TIMIT

Aquí se describirán las características principales del corpus elegido de manera de definir perfectamente los alcances y complejidad de la tarea de clasificación de fonemas extraídos del mismo, para más detalles remitirse a la documentación suministrada con la base de datos [GLF93]. Esta base de datos ha sido confeccionada en forma conjunta por Texas Instruments (TI) y el Massachusetts Institute of Technology (MIT). Consiste en una serie de emisiones de voz grabadas a través de la lectura de diversos textos por un conjunto de hablantes. Esta base ha sido diseñada para la adquisición de conocimiento acústico-fonético a partir de los datos de voz y para el desarrollo y evaluación de sistemas de RAH. TIMIT contiene la voz de 630 hablantes representando las 8 mayores divisiones dialécticas del Inglés Americano, cada uno pronunciando 10 oraciones fonéticamente ricas. El corpus TIMIT incluye la señal de voz correspondiente a cada oración hablada, así como también transcripciones ortográficas, fonéticas y de palabras alineadas temporalmente. Además los datos vienen ya divididos en subconjuntos de entrenamiento y prueba balanceados para cobertura dialéctica y fonética lo que facilita también la comparación de resultados. La versión de la base de datos empleada en el trabajo es la 1-1.1 de Octubre de 1990.

TIMIT contiene un total de 6300 oraciones, 70 % de los hablantes son masculinos y 30 % son femeninos. El material de texto consiste de 2 *oraciones de dialecto* (SA), 450 *oraciones fonémicamente compactas* (SX), y 1890 *oraciones fonéticamente diversas* (SI). Cada hablante lee las 2 SA, 5 de las SX y 3 de las SI.

Organización de los datos

El CD-ROM contiene una estructura de árbol de directorios jerárquica que permite fácil acceso a los datos en forma automática (por medio del programa de procesamiento). La estructura de este árbol es la siguiente :

```
/<Corpus>/<Uso>/<Dialecto>/<Sexo><Hablante>/<Oración>.<Tipo_Archivo>
donde,
CORPUS = timit
USO = train|test (entrenamiento|prueba)
DIALECTO = dr1|dr2|dr3|dr4|dr5 dr6|dr7|dr8 (regiones dialécticas)
SEXO = f|m
HABLANTE = <INICIALES><DÍGITO>
donde,
INICIALES = Iniciales del Hablante (3 letras)
DÍGITO = número 0-9 para diferenciar hablantes iniciales iguales
ORACIÓN = <TIPO_TEXTO><NÚMERO_ORACIÓN>
donde,
TIPO_TEXTO = sa|si|sx
NÚMERO_ORACIÓN = 1 ... 2342

TIPO_ARCHIVO = wav|txt|wrd|phn (datos de voz|texto|palabras|fonemas )
```

Por ejemplo : "/timit/train/dr1/fcjf0/sa1.wav", corresponde al corpus TIMIT, conjunto de entrenamiento, región dialéctica 1, sexo femenino, hablante "cjf0", texto oración "sa1", archivo señal de voz.

Tipos de Archivo

TIMIT incluye varios archivos asociados con cada emisión (Tabla 1). Así mismo se incluye un archivo diccionario con todas las palabras contenidas en el corpus y su transcripción fonética (léxico), otro con todas las oraciones empleadas, y otro con información específica de los hablantes. Los archivos de voz poseen el formato con cabecera NIST SPHERE que ha sido diseñado para facilitar el intercambio de datos de señales de voz en CD-ROM. La cabecera NIST es una estructura orientada a objetos de 1024 bytes que precede a los datos propiamente dichos. En ella se almacena información acerca de la emisión como ser frecuencia de muestreo, bits por muestra, identificación del hablante y la oración, etc.

TIPO ARCHIVO	DESCRIPCIÓN
.wav	Archivo de voz con cabecera tipo SPHERE.
.txt	Transcripción ortográfica asociada de las palabras dichas por el hablante.
.wrd	Transcripción de palabras alineada temporalmente con el archivo de voz.
.phn	Transcripción fonética alineada temporalmente con el archivo de voz.

Tabla 1: Tipos de Archivos asociados a cada emisión.

Los archivos de transcripción tienen la siguiente forma :

```
<MUESTRA_COMIENZO><MUESTRA_FINAL> <TEXTO><nueva-línea>
```

```
<MUESTRA_COMIENZO><MUESTRA_FINAL> <TEXTO><nueva-línea>
.
.
.
<MUESTRA_COMIENZO><MUESTRA_FINAL> <TEXTO><nueva-línea>
donde,
MUESTRA_COMIENZO = Muestra inicial del segmento (número entero >=0)
MUESTRA_FINAL = Muestra final del segmento (número entero <= última
muestra)
TEXTO = <ORTOGRAFÍA> | <ETIQUETA_PALABRA> | <ETIQUETA_FONÉTICA>
donde,
      ORTOGRAFÍA = Transcripción ortográfica completa del texto.
      ETIQUETA_PALABRA = Una palabra de la transcripción ortografía.
      ETIQUETA_FONÉTICA = Un código de transcripción fonética.
```

Por ejemplo las transcripciones de la emisión en "timit\test\dr1\mjsw0\si1640.wav" serían :

Ortografía (.txt):

```
0 26112 How did one join them?
```

Palabras (.wrd) :

```
2276 5111 how
5111 8003 did
8003 12560 one
12560 19174 join
19174 22747 them
```

Fonética (.phn) :

```
0 2276 h#
2276 3320 hh
3320 5111 aw
5111 5931 dcl
5931 6143 d
6143 7272 ih
7272 8003 dcl
8003 9869 w
9869 11224 ah
11224 12560 n
12560 12879 dcl
12879 14160 jh
14160 17560 oy
17560 19174 n
19174 19533 dh
19533 21400 eh
21400 22747 m
22747 26000 h#
```

En la Figura 17 se puede apreciar la emisión correspondiente con la superposición de las etiquetas de palabras y fonemas alineadas temporalmente con la misma.

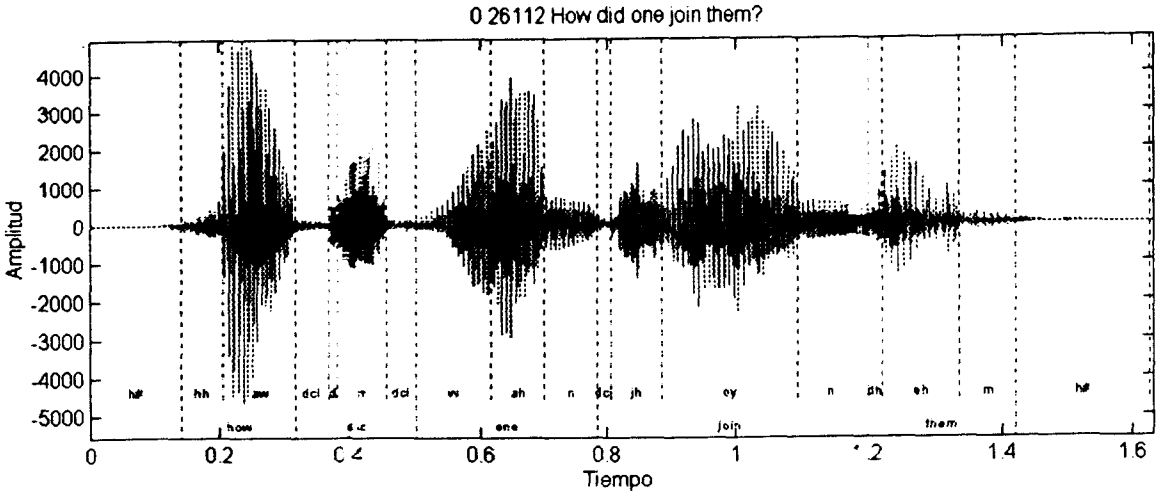

Figura 17: Señal de voz con etiquetas de palabras y fonemas

Selección de Hablantes

Las 10 oraciones leídas por cada uno de los 630 hablantes representan aproximadamente 30 segundos de voz por hablante. En total el corpus contiene aproximadamente 5 horas de voz. Todos los participantes seleccionados fueron hablantes nativos de Inglés Americano. Además todos fueron calificados como sin patologías clínicas del habla por un especialista del área. Se detectaron en algunos sujetos pequeñas anormalidades en el habla o la audición que fueron anotadas en los archivos de información de los hablantes que acompañan la base de datos. Los hablantes fueron seleccionados para ser representativos de diferentes regiones dialécticas geográficas de los Estados Unidos de acuerdo con la región donde vivieron en su niñez. En la Tabla 2 se presenta la distribución de los hablantes en cada región dialéctica.

Condiciones de Grabación

Las grabaciones fueron hechas en una cabina de grabación aislada de ruidos usando un sistema semiautomático para la presentación del texto al hablante y la grabación. Los datos fueron digitalizados a una frecuencia de muestreo de 20 KHz (16 bits) con un filtro anti-alias en 10 KHz. La voz fue filtrada digitalmente, nivelada (debiased) y submuestreada a 16 KHz [FWD86]. A los sujetos se los estimuló con una señal de ruido de fondo de bajo nivel a través de auriculares para suprimir la inusual calidad de voz producida por el efecto de aislación de la cabina. También se les pidió que leyeran el texto con "voz natural".

Texto del Corpus

Las oraciones SA fueron diseñadas para exponer las diferencias dialécticas y fueron leídas por todos los hablantes. Las oraciones SX fueron diseñadas a mano para proveer una buena cobertura en cuanto a pares de fonemas, con ocurrencias extra de contextos fonéticos difíciles o de interés particular. Cada hablante leyó 5 de estas oraciones y cada una fue leída por 7 hablantes. Las oraciones SI fueron seleccionadas de fuentes de texto existentes para agregar diversidad en los tipos de oraciones y los contextos fonéticos. El criterio de

selección maximiza la variedad de contextos alofónicos encontrados en los textos. Cada hablante leyó 3 de estas oraciones y cada una fue leída solo una vez. En la Tabla 3 se muestra la distribución del material de texto del corpus.

Región Dialéctica		Nº Hablantes	Nº Hablantes	Nº Total de
Nombre	Código	Masculinos	Femeninos	Hablantes
New England	1	31 (63%)	18 (27%)	49 (8%)
Northern	2	71 (70%)	31 (30%)	102 (16%)
North Midland	3	79 (67%)	23 (23%)	102 (16%)
South Midland	4	69 (69%)	31 (31%)	100 (16%)
Southern	5	62 (63%)	36 (37%)	98 (16%)
New York City	6	30 (65%)	16 (35%)	46 (7%)
Western	7	74 (74%)	26 (26%)	100 (16%)
Army Brat	8	22 (67%)	11 (33%)	33 (5%)
Nº Total de Hablantes		438 (70%)	192 (30%)	630 (100%)

Tabla 2: Distribución de los Hablantes

Subdivisión en Entrenamiento y Prueba

Existen diferentes métodos para estimar la capacidad de generalización de un clasificador [MST94]. Es ampliamente conocido que las tasas de error tienden a sesgarse si se estiman a partir de los mismos datos que se utilizaron en el proceso de aprendizaje o entrenamiento del clasificador. Una forma muy sencilla (y difundida) de abordar el problema consiste en separar los datos en un conjunto de entrenamiento y otro de prueba. En la sección sobre el clasificador se ampliará más este punto. Sin embargo, aquí es importante notar que la cantidad de datos involucrados en este problema hace imposible la utilización de métodos más precisos como Validación Cruzada o Bootstrap para estimar el error.

El material contenido en TIMIT fue dividido en conjuntos de entrenamiento y prueba siguiendo los siguientes criterios :

1. Del 20 al 30 % del corpus sería usado para propósitos de prueba dejando el restando 70 a 80 % para entrenamiento.

2. Ningún hablante debería aparecer en ambos conjuntos.

3. Todas las regiones dialécticas deberían estar representadas en ambos conjuntos, con al menos un hablante masculino y uno femenino de cada dialecto.

4. La cantidad de material de texto repetido en ambos conjuntos debería minimizarse o, en lo posible, eliminarse.

5. Todos los fonemas deberían estar cubiertos en el material de prueba, preferiblemente en diferentes contextos.

Estos criterios, junto con lo que se mencionó en la introducción, hacen que el problema de reconocimiento o clasificación sea independiente del hablante y del texto, lo que implica un grado de complejidad apreciable teniendo en cuenta la cantidad de material disponible.

Tipo Oración	Nº Oraciones	Nº Hablantes/Oración	Total	Nº Oraciones/Hablante
Dialecto (SA)	2	630	1260	2
Compactas (SX)	450	7	3150	5
Diversas (SI)	1890	1	1890	3
Total	2342		6300	10

Tabla 3: Material de texto TIMIT

Códigos de Símbolos Fonémicos y Fonéticos

Aquí presentaremos las tablas con los símbolos fonémicos y fonéticos usados en el léxico de TIMIT y en las transcripciones fonéticas.. Estos incluyen marcadores de intensidad (stress) {1,2} encontrados solo en el léxico y los siguientes símbolos que ocurren solo en las transcripciones:

1. Los intervalos de cierre u oclusión de las oclusivas los cuales se distinguen de la liberación o explosión de las mismas. Los símbolos de la oclusión para /b/, /d/, /g/, /p/, /t/, /k/ son /bcl/, /dcl/, /gcl/, /pcl/. /tck/, /kcl/, respectivamente. Las porciones de oclusión de /jh/ y /ch/, son /dcl/ y /tcl/.

2. Alófonos que no ocurren en el léxico. El uso de determinado alófono puede depender del hablante, del dialecto, la velocidad de emisión y el contexto fonémico entre otros factores. Dado que el uso de estos alófonos es difícil de predecir no han sido usados en las transcripciones fonéticas del léxico.

3. Otros símbolos incluyen dos tipos de silencio, pau indicando una pausa, y epi, denotando el silencio epentético que es frecuentemente encontrado entre una fricativa y una semivocal o nasal, además de h#, usado para marcar el silencio y/o no aparición de eventos de voz encontrado al principio o al final de la señal.

TIPO	SÍMBOLO	PALABRA EJEMPLO	TRANSCRIPCIÓN FONÉTICA
Oclusivas	b	bee	BCL B iy
	d	day	DCL D ey
	g	gay	GCL G ey
	p	pea	PCL P iy
	t	tea	TCL T iy
	k	key	KCL K iy
	dx	muddy, dirty	m ah DX iy, dcl d er DX iy
	q	bat	bcl b ae Q
Africadas	jh	joke	DCL JH ow kcl k
	ch	choke	TCL CH ow kcl k
Fricativas	s	sea	S iy
	sh	she	SH iy
	z	zone	Z ow n
	zh	azure	ae ZH er
	f	fin	F ih n
	th	thin	TH ih n
	v	van	V ae n
	dh	then	DH e n
Nasales	m	mom	M aa M
	n	noon	N uw N
	ng	sing	s ih NG
	em	bottom	b aa tcl t EM
	en	button	b ah q EN
	eng	washington	w aa sh ENG tcl t ax n
	nx	winner	w ih NX axr
Semivocales y Glides	l	lay	L ey
	r	ray	R ey
	w	way	W ey
	y	yacht	Y aa tcl t
	hh	hay	HH ey
	hv	ahead	ax HV eh dcl d
	el	bottle	bcl b aa tcl t EL
Vocales	iy	beet	bcl b IY tcl t
	ih	bit	bcl b IH tcl t
	eh	bet	bcl b EH tcl t
	ey	bait	bcl b EY tcl t
	ae	bat	bcl b AE tcl t
	aa	bott	bcl b AA tcl t
	aw	bout	bcl b AW tcl t
	ay	bite	bcl b AY tcl t
	ah	but	bcl b AH tcl t
	ao	bought	bcl b AO tcl t
	yo	boy	bcl b YO
	ow	boat	bcl b OW tcl t
	uh	book	bcl b UH kcl k
	uw	boot	bcl b UW tcl t
	ux	toot	tcl t UX tcl t
	er	bird	bcl b ER dcl d
	ax	about	AX bcl b aw tcl t
	ix	debit	dcl d eh bcl b IX tcl t
	axr	butter	bcl b ah dx AXR
	ax-h	suspect	s AX-H s pcl p eh kcl k tcl t

Tabla 4: Símbolos fonéticos utilizados en la transcripción

TIPO	SÍMBOLO	DESCRIPCIÓN
Otros	pau	pausa
	epi	silencio epentético
	h#	marcador de comienzo / fin
	1	marcador de stress primario
	2	marcador de stress secundario

Tabla 5: Otros símbolos empleados

La cantidad total de símbolos que se pueden utilizar en la clasificación es de 52. Estos se distribuyen como 8 tipos de fonemas oclusivos, 2 africados, 15 fricativos, 7 semivocales y glides y 20 vocales.

Datos elegidos para los experimentos

Como se puede apreciar la cantidad de símbolos o fonemas a clasificar y la de emisiones y hablantes es demasiado grande para intentar realizar los experimentos con toda la base de datos (que además totaliza unos 650 MBytes de información). Por esta razón se debieron establecer algunos criterios para utilizar un subconjunto menor de los fonemas y hablantes de la base y sin embargo poder extrapolar los resultados a todo el conjunto. Estos criterios fueron :

1. Utilizar un subconjunto de fonemas de relativa dificultad de diferenciación.

2. Cubrir los tipos de fonemas más importantes.

3. Disminuir la cantidad de hablantes y la diversidad de dialectos.

Se sabe por experimentos psico-acústicos que las consonantes /b/ y /d/ del tipo oclusivo son difíciles de distinguir en varios contextos. Por otra parte el fonema /jh/ es africado con lo que se incluirían las características especiales de este grupo (que posee un componente oclusivo seguido de uno fricativo). Además, estas son algunas de las consonantes iniciales del conocido E-Set de TI-46 que ha probado también ser un subconjunto de palabras difícil de clasificar por medios automáticos. Para agregar a este grupo algunas vocales se eligieron /eh/ e /ih/ cuya distancia en el espacio de formantes es muy pequeña. Esto las convierte en otro grupo altamente confundible. En la Figura 18 se puede observar la distribución de las vocales del Inglés en función de f1 y f2 (los datos usados para construir la gráfica fueron tomados de [PeB52]). De esta manera nuestro subconjunto está formado por 5 fonemas (10 % del total).

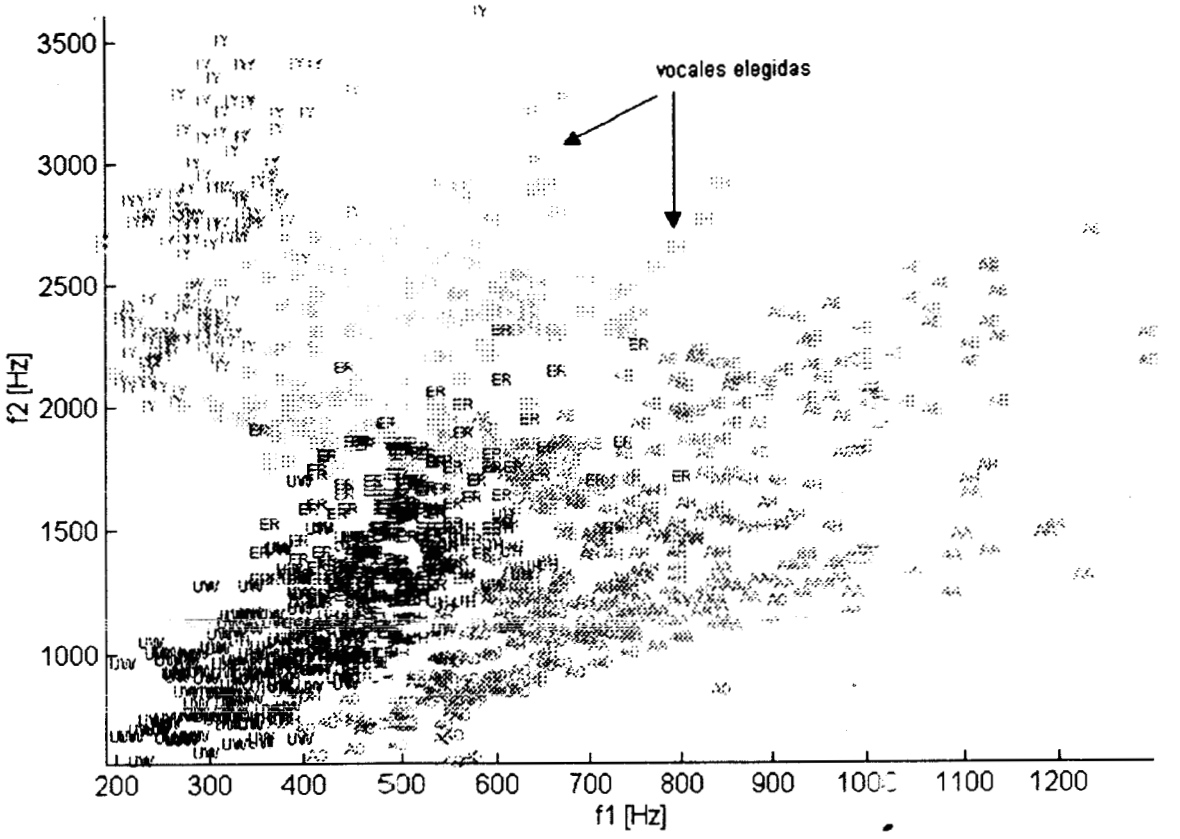

Figura 18: Distribución de las vocales del Inglés

Siguiendo los criterios expuestos se eligió la región "dr1" que posee casi 50 hablantes (ver Tabla 2). Se respetaron las divisiones en conjuntos de entrenamiento y prueba propuesta en TIMIT de manera que la distribución final de los fonemas elegidos en cada región se puede apreciar en la Tabla 6.

Fonema	Entrenamiento	Prueba	Total
/b/	183 (14.4 %)	59 (15.9 %)	242
/d/	300 (23.6 %)	90 (24.2 %)	390
/jh/	104 (8.2 %)	20 (5.3 %)	124
/eh/	316 (24.8%)	93 (25.1 %)	409
/ih/	370 (29.0 %)	109 (29.4 %)	479
Total	1273	371	1644

Tabla 6: Distribución de los fonemas elegidos en entrenamiento y prueba.

Se debe aclarar que en primera instancia se incluyó en los archivos la oclusión de /b/, /d/ y /jh/, pero debido a que para estas dos últimas los símbolos asociados (y también el fenómeno acústico) resultan idénticos era imposible diferenciarlas. Por esta razón se excluyeron de los experimentos definitivos.

Fonema	/b/	/d/	/jh/	/eh/	/ih/
Muestras	300	378	916	1419	1218
Tiempo	19 ms	24 ms	57 ms	89 ms	76 ms
Frames	2	3	7	11	10

Tabla 7: Duración promedio de los fonemas.

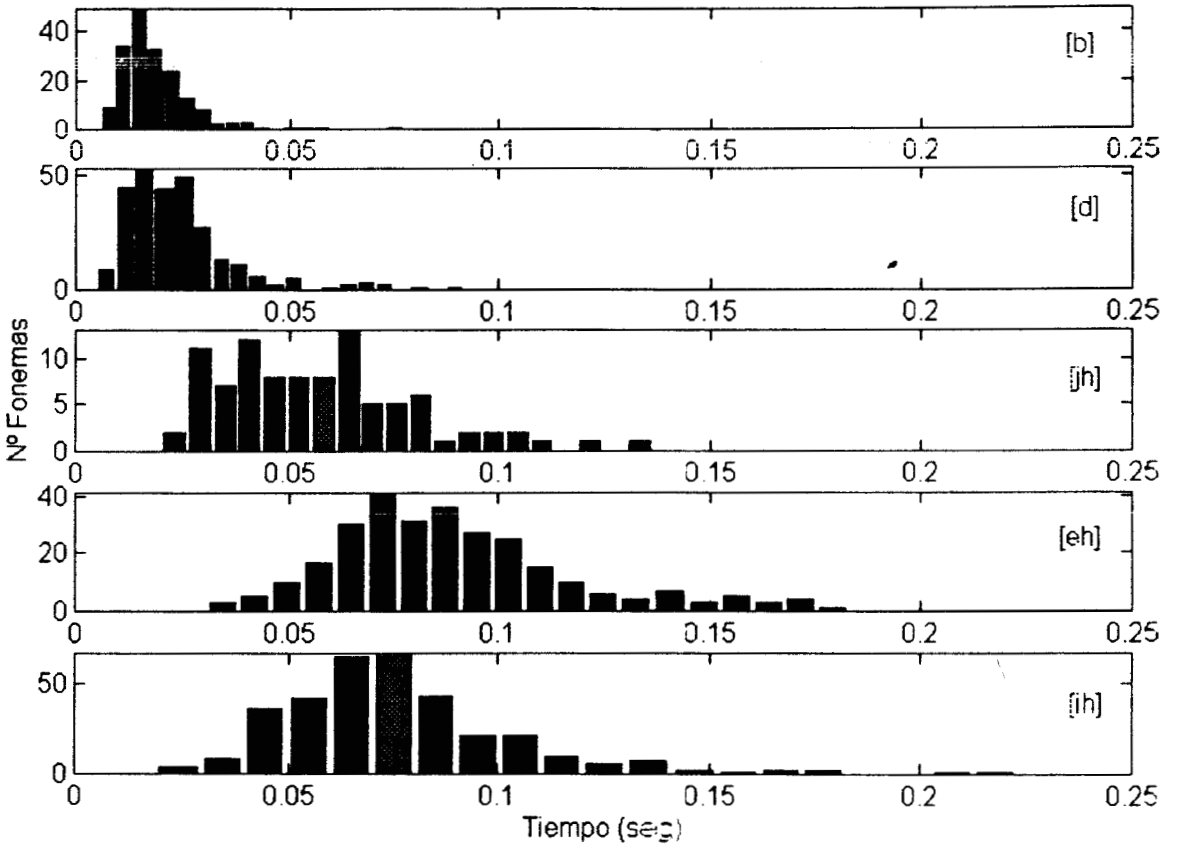

Figura 19: Histogramas de duración de los fonemas

Como nos interesa seguir la evolución temporal de distintas características de la señal se utiliza un esquema que aplica el procesamiento (Fourier o Wavelets) a una ventana de la señal. El ancho de la ventana es del orden de los 10 mseg y cada patrón generado se denomina *frame*. Este tema se ampliará en la sección sobre procesamiento. También se debe señalar la gran variación de duración de los distintos fonemas. Por ejemplo /b/ y /d/ son

Se debe aclarar que en primera instancia se incluyó en los archivos la oclusión de /b/, /d/ y /jh/, pero debido a que para estas dos últimas los símbolos asociados (y también el fenómeno acústico) resultan idénticos era imposible diferenciarlas. Por esta razón se excluyeron de los experimentos definitivos.

Fonema	/b/	/d/	/jh/	/eh/	/ih/
Muestras	300	378	916	1419	1218
Tiempo	19 ms	24 ms	57 ms	89 ms	76 ms
Frames	2	3	7	11	10

Tabla 7: Duración promedio de los fonemas.

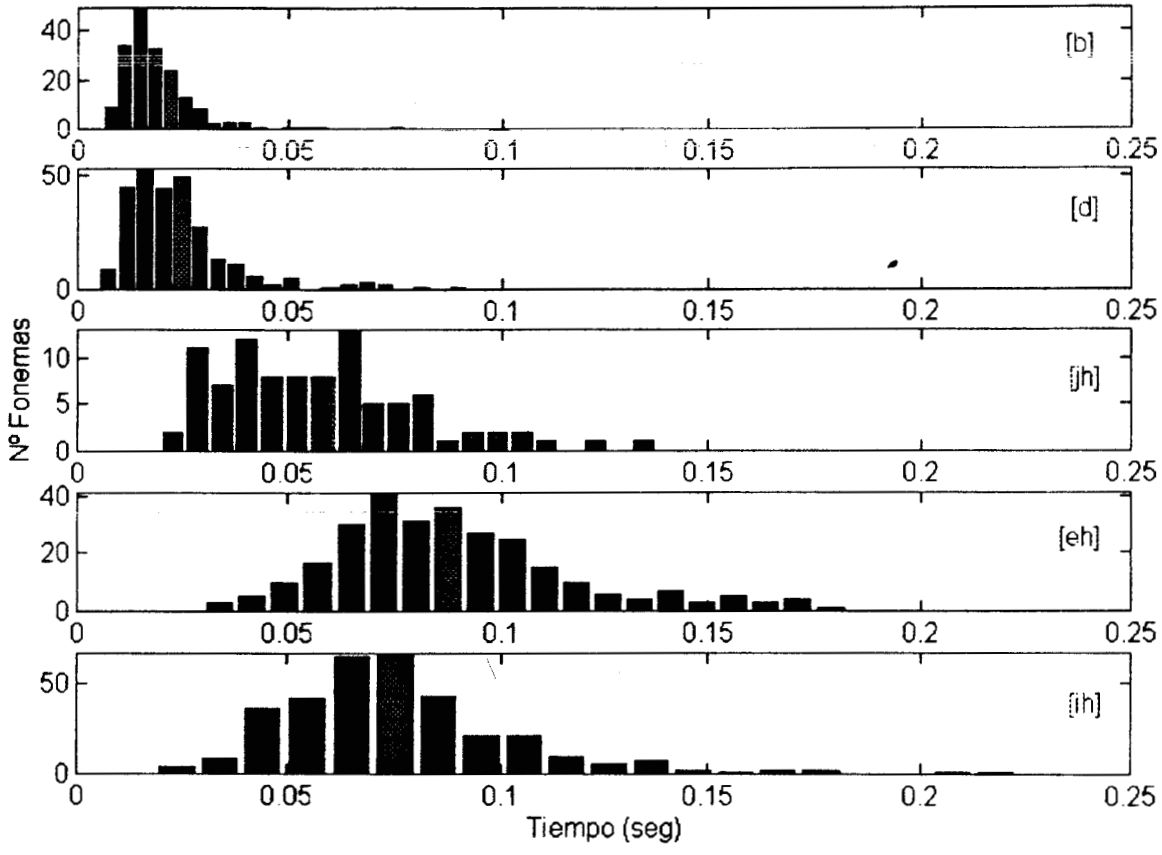

Figura 19: Histogramas de duración de los fonemas

Como nos interesa seguir la evolución temporal de distintas características de la señal se utiliza un esquema que aplica el procesamiento (Fourier o Wavelets) a una ventana de la señal. El ancho de la ventana es del orden de los 10 mseg y cada patrón generado se denomina *frame*. Este tema se ampliará en la sección sobre procesamiento. También se debe señalar la gran variación de duración de los distintos fonemas. Por ejemplo /b/ y /d/ son

generalmente muy cortas frente a /eh/, /ih/ o incluso /jh/ (ver Tabla 7 y Figura 19), esto junto con sus respectivas distribuciones (Tabla 6) lleva a que una vez procesadas la cantidad de frames o patrones generados para cada clase sea muy diferente, produciendo algunos problemas sobre las clases menos representadas . Por otra parte la diferencia en duración también plantea problemas en cuanto a los procesos involucrados en la clasificación dinámica de los patrones.

En las figuras siguientes se pueden apreciar algunos sonogramas como ejemplo de los fonemas elegidos para la hablante FCJF0 de la región DR1. En la Figura 20 vemos el caso de una /b/ en SI1657 y en la Figura 21 otra 'b/ de la misma emisión. Obsérvense las grandes diferencias a nivel temporal y espectral (hay que tener en cuenta que corresponden al mismo hablante y a dos instantes de la misma emisión).

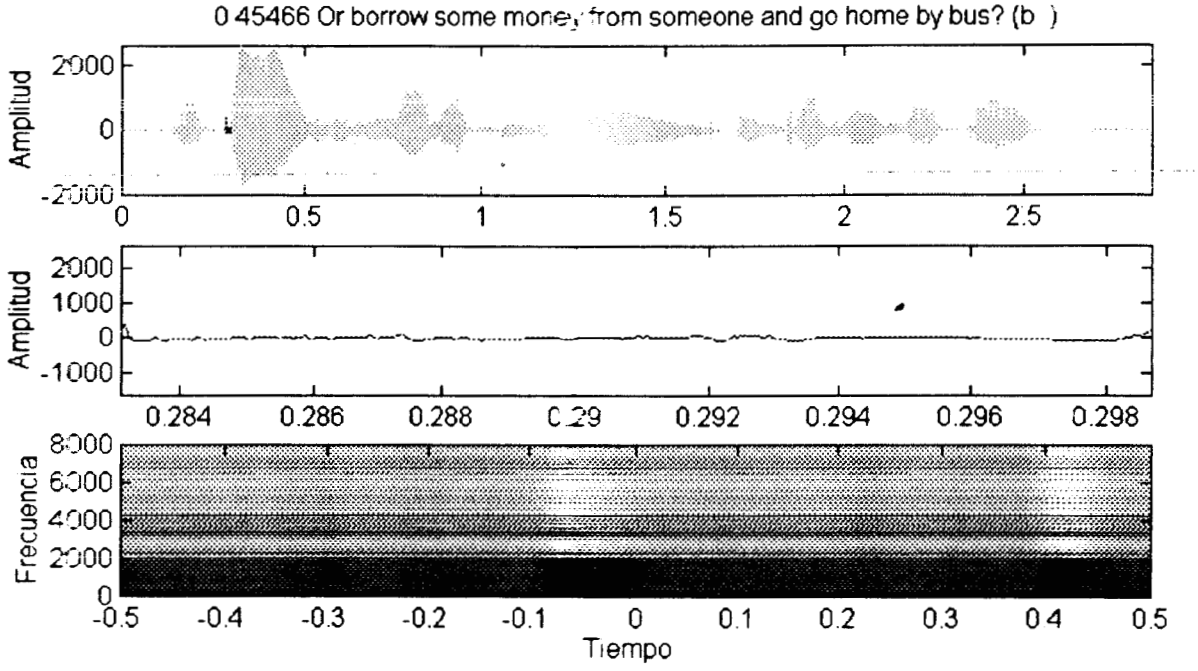

Figura 20 : Fonema /b/ en SI1657

En la Figura 22 vemos una /d/ en SA1, se puede apreciar la corta duración (similar a la de la /b/) y los patrones regulares en el espectro que delatan su componente glótica. A continuación (Figura 23) se presentan idénticos análisis de la /jh/ donde se distingue el contenido de alta frecuencia debido a la componente fricativa. En la Figura 24 y la Figura 25 se muestran las gráficas de dos /eh/, una en SA1 y la otra en SI1657, donde se puede apreciar las diferencias en amplitud y duración de ambas realizaciones del fonema. Por último en la Figura 26 se observa la vocal /ih/ en SA1.

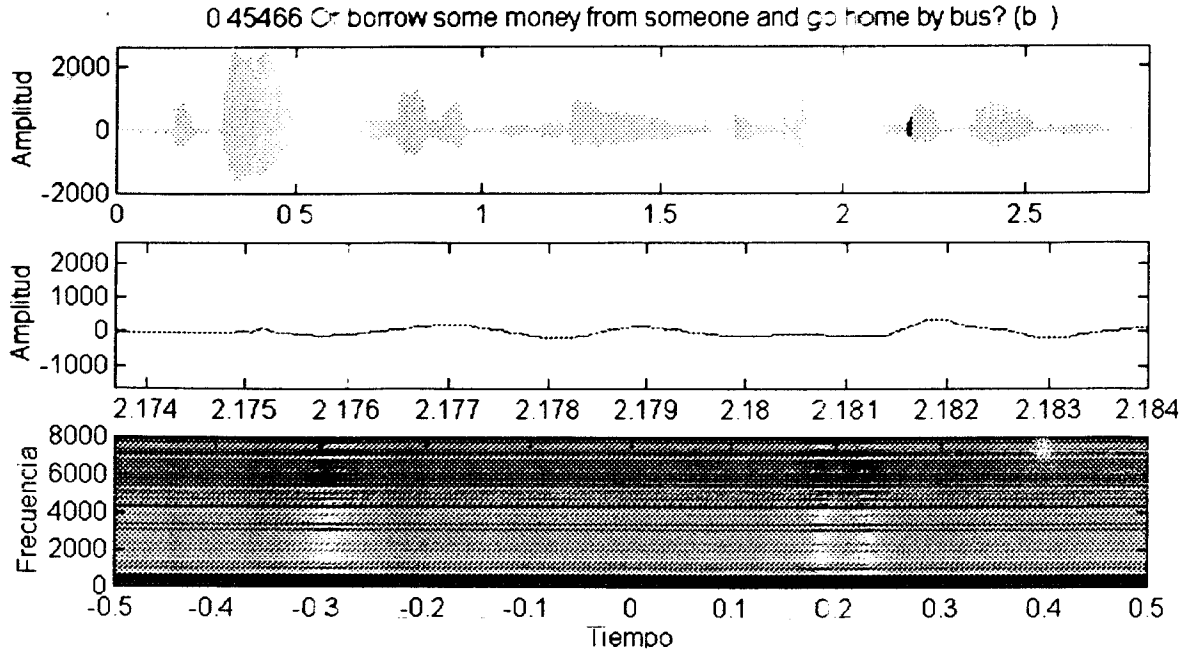

Figura 21 : Otra /b/ en SI1657

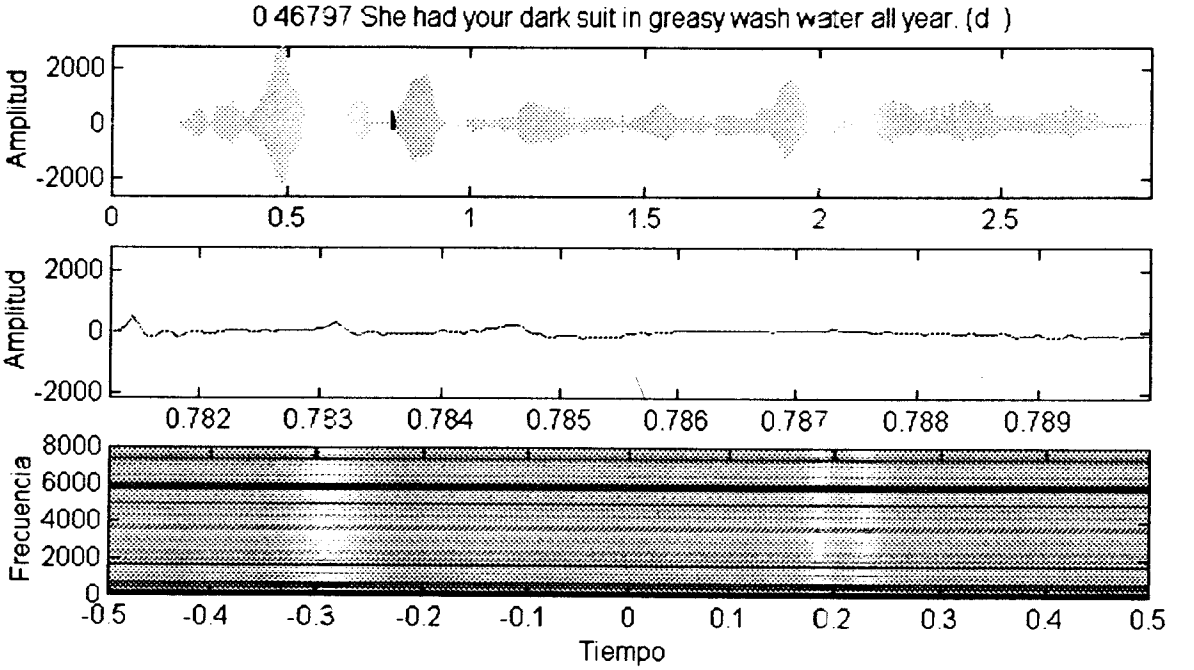

Figura 22: Fonema /d/ en SA1

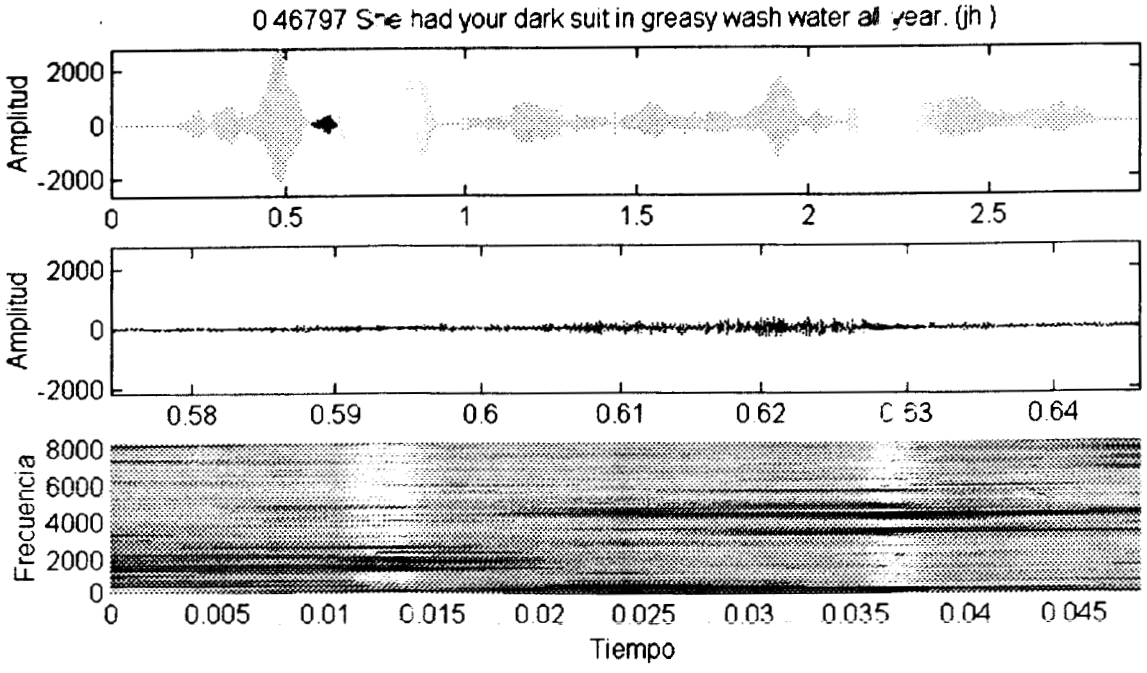

Figura 23: Fonema /jh/ en SA1.

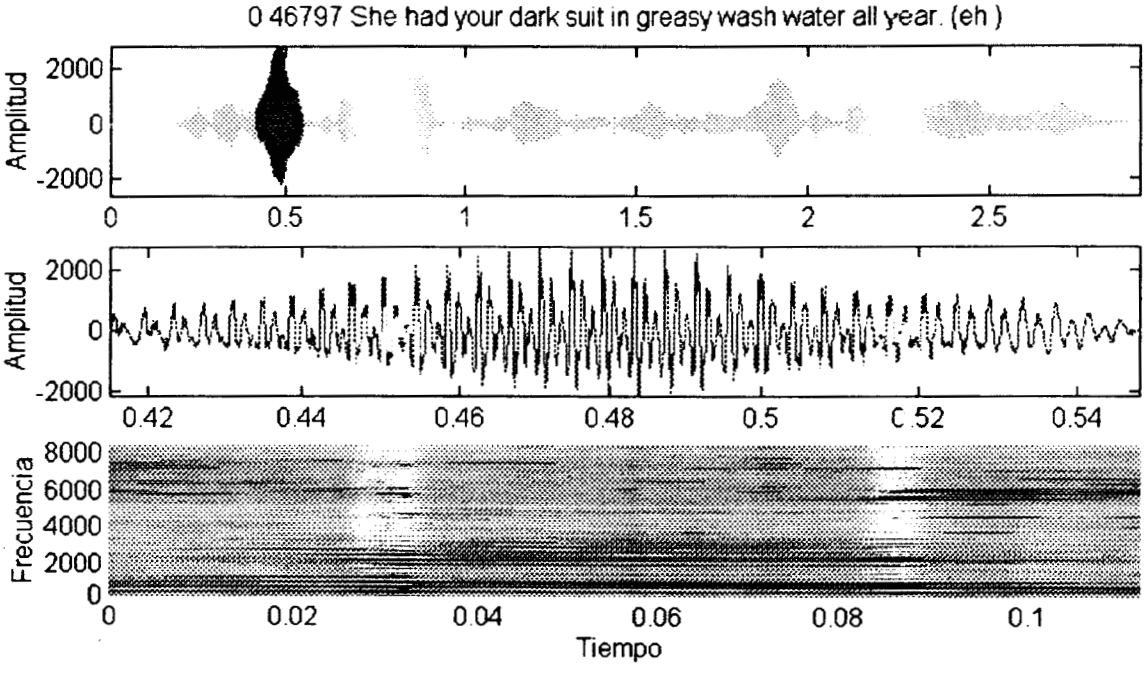

Figura 24: Fonema /eh/ en SA1.

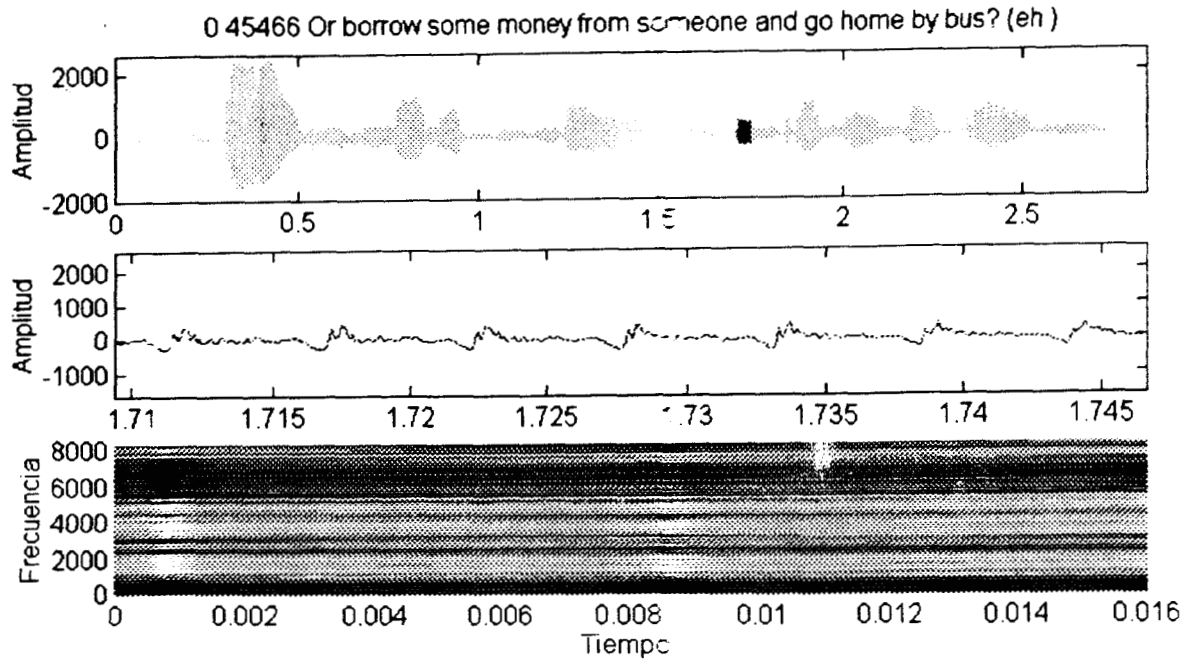

Figura 25: Fonema /eh/ en SI1657

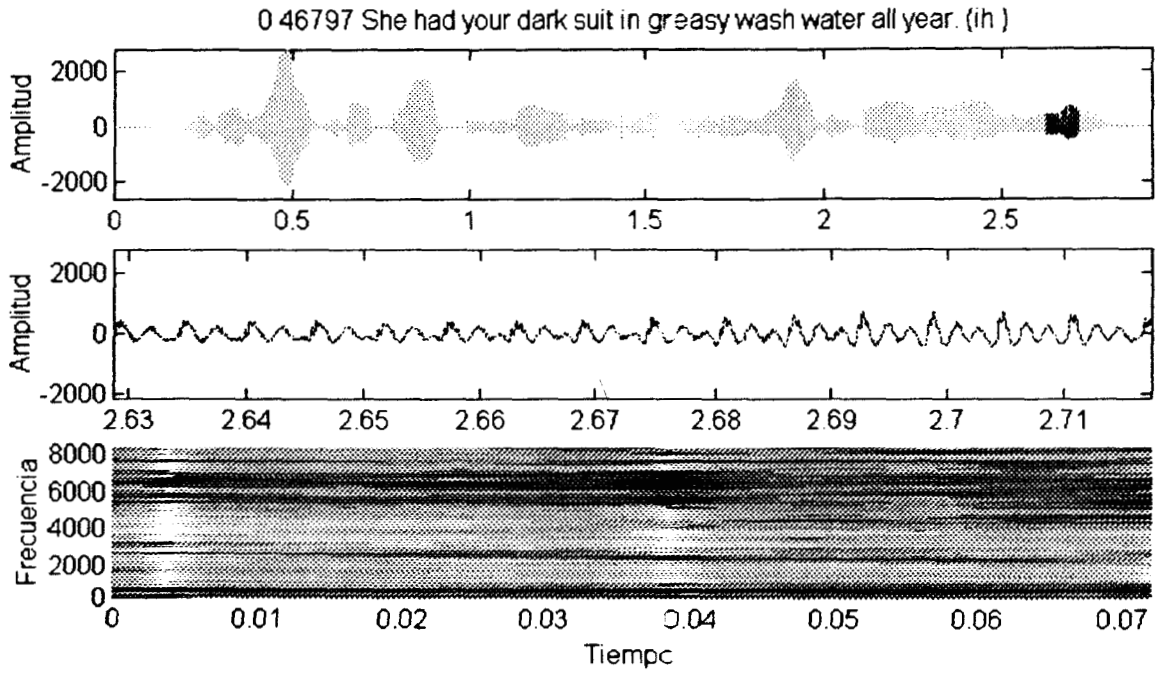

Figura 26: Fonema /ih/ en SA1.

IV . El Procesamiento

Introducción

Como ya se mencionó en la introducción general el tipo de análisis clásico para las señales de voz hasta la actualidad ha sido la Transformada de Fourier de Tiempo Corto (STFT). Se puede decir que la Transformada de Fourier (FT) ha dominado el campo de análisis de señales por mucho tiempo [Fou88]. Sin embargo recientemente se ha desarrollado la Transformada Wavelet (WT) que permite realizar el análisis de señales no estacionarias en forma más "eficiente".

La STFT posee limitaciones para el análisis de señales con componentes transitorias debidas a su resolución fija. La WT constituye una alternativa a esta técnica. La principal diferencia es que, en contraste con la STFT, que usa una ventana de análisis única, la WT usa ventanas pequeñas a altas frecuencias y ventanas grandes a bajas frecuencias. Este comportamiento es similar al análisis que realiza el oído según se estudió en el capítulo de aspectos fisiológicos.

En caso de las señales digitales la representación o Transformada Wavelet Discreta (DWT) tiene varias características atractivas que han contribuido a su reciente aumento de popularidad en ámbitos de matemáticas y procesamiento de señales:

1. Su descomposición jerárquica que permite la caracterización de la señal a distintas escalas (análisis multiresolución).

2. Es en esencia una descomposición de la señal en sub-bandas; lo que está muy relacionado con una gran variedad de técnicas de descomposición multifrecuencia.

3. Existe un algoritmo rápido para calcularla de implementación digital sencilla.

Todas estas características y una serie de aplicaciones recientes exitosas de esta técnica a nuestro campo [FaK93], [FaK94], [WeW92], [Wic91] nos han llevado a considerarla como una alternativa interesante a los métodos clásicos para análisis de voz.

Este capítulo se organizará de la siguiente manera. Primero se presenta la transformada de Fourier y su limitación para tratar con señales transitorias. Esto nos lleva directamente a la Transformada de Fourier de Tiempo Corto. Una vez mostrados sus inconvenientes se procede a introducir la Transformada Wavelet. Luego se exponen los principios del Análisis Multiresolución en los que se basa la transformada. A continuación se describen las familias más importantes de Wavelets. Luego se exponen algunas consideraciones relacionadas con la elección del análisis adecuado para nuestro caso. Por último se describe lo concerniente a la implementación práctica de los algoritmos.

Transformada de Fourier

En lo que sigue se presentarán los aspectos conceptuales esenciales de la FT que nos permitirán comprender mejor la WT, para una excelente revisión del tema ver [RiV91].

El objetivo del procesamiento de señales es el de extraer la información relevante de una señal por medio de algún tipo de transformación. Algunos métodos realizan suposiciones *a priori* acerca de la señal que se va a analizar. Por ejemplo los métodos paramétricos, como los basados en modelos ARMA, suponen que el sistema que generó la señal es de este tipo (o sea un sistema ARMA). Esto lleva a excelentes resultados si estas suposiciones son válidas, pero obviamente no es de aplicabilidad general.

Estas transformaciones han sido aplicadas a señales estacionarias, es decir, aquellas cuyas propiedades no cambian con el tiempo. Para esta clase de señales $x(t)$, la transformación estacionaria más "natural" es la conocida FT [Fou88]:

$$X(f) = \int_{-\infty}^{\infty} x(t) \cdot e^{-2\pi i f t} dt \qquad (1)$$

Los coeficientes de análisis $X(f)$ definen la noción de frecuencia global en una señal. Como se muestra en (1), estos se calculan como si fueran productos internos (aunque en este caso las integrales van desde -∞ a ∞) de la señal con las funciones senoidales de duración infinita de la base. Como resultado, el análisis de Fourier funciona bien si $x(t)$ esta compuesta de un número reducido de componentes estacionarias. Sin embargo, cualquier cambio abrupto en una señal no estacionaria $x(t)$ se esparce sobre todo el eje de frecuencias en $X(f)$. Es por ello que para el correcto análisis de señales no estacionarias se requiere más que la transformada de Fourier.

La aproximación más común para resolver este problema consiste en introducir la dependencia temporal en el análisis de Fourier preservando su linealidad. La idea es introducir una parámetro de "frecuencia local" (local en el tiempo) de tal forma que la transformada de Fourier local mire a la señal a través de una ventana sobre la cual esta es aproximadamente estacionaria. Otra forma, equivalente, es modificar las funciones senoidales de la base de manera que se concentren más en el tiempo (y como consecuencia menos en la frecuencia).

Transformada de Fourier de Tiempo Corto

La "frecuencia instantánea" [Fla89] ha sido considerada frecuentemente como una forma de introducir la dependencia del tiempo. Sin embargo, si la señal no es de banda angosta, la frecuencia instantánea promedia diferentes componentes espectrales en el tiempo. Para volverse precisa en el tiempo se necesita una representación tiempo-frecuencia (que denominaremos $S(t, f)$ por su dependencia de t y f) de la señal $x(t)$ compuesta de características espectrales dependientes del tiempo, definiendo la frecuencia local f de una manera adecuada a través de $S(t, f)$. Esta representación es similar a la notación usada en música, la cual muestra también "frecuencias" (notas) tocadas en el tiempo.

La transformada de Fourier (1), fue adaptada por primera vez por Gabor [Gab46] para definir $S(t, f)$ como sigue. Considere una señal $x(t)$ y asuma que es estacionaria si se la observa a través de una ventana $g(t)$ de extensión limitada, centrada en el tiempo τ. La transformada de Fourier (1) de las señales ventaneadas $x(t).g^*(t-\tau)$ (donde $g^*(t)$ representa el conjugado de $g(t)$ en el caso complejo) nos lleva a la transformada de Fourier de Tiempo Corto (STFT):

$$STFT(\tau, f) = \int x(t) \cdot g^*(t-\tau) \cdot e^{-2j\pi ft} dt \qquad (2)$$

que mapea la señal en una función bidimensional en el plano tiempo-frecuencia (τ, f).

El parámetro f en (2) es similar a la frecuencia de Fourier y esta transformación hereda varias de las propiedades de la transformada de Fourier. Sin embargo, aquí el análisis depende de la elección de la ventana $g(t)$. Este punto de vista muestra a la STFT como un proceso de ventaneo de la señal.

Una visión alternativa está basada en la interpretación del mismo proceso como un banco de filtros. Para una frecuencia f dada, (2) filtra la señal en el tiempo con un filtro pasa-banda cuya respuesta al impulso es la función ventana modulada a esa frecuencia. De esta manera la STFT puede ser vista como un banco de filtros modulado [All77],[Por80].

De esta doble interpretación se pueden derivar algunas relaciones respecto a la resolución de la transformada en el tiempo y en la frecuencia. Dada una función ventana $g(t)$ y su transformada de Fourier $G(f)$, se define el ancho de banda Δf del filtro como:

$$\Delta f^2 = \frac{\int f^2 \cdot |G(f)|^2 df}{\int |G(f)|^2 df} \qquad (3)$$

Dos sinusoides pueden discriminarse solo si están más separadas que Δf, por lo que define la resolución en frecuencia de la STFT. De forma similar la dispersión en el tiempo está dada por :

$$\Delta t^2 = \frac{\int t^2 \cdot |g(t)|^2 dt}{\int |g(t)|^2 dt} \quad (4)$$

Dos pulsos pueden discriminarse solo si están más lejos que Δt. Ahora ni la resolución temporal, ni la frecuencial pueden ser arbitrariamente pequeñas, porque su producto debe cumplir la siguiente relación conocida como **principio de incertidumbre**:

$$\Delta t \cdot \Delta f \geq \frac{1}{4\pi} \quad (5)$$

Una cuestión fundamental con respecto a la STFT es que una vez que se elige una ventana la resolución tiempo-frecuencia queda fija para todo el análisis. Esto ocasiona que si por ejemplo se quiere analizar una señal compuesta de pequeños transitorios junto con componentes cuasi-estacionarias esta puede ser analizada con buena resolución en tiempo o en frecuencia, pero no ambas. Esto ocurre frecuentemente con señales de voz y es lo que ha llevado a la utilización conjunta de dos tipos de espectrogramas para analizar las distintas características de la voz (Figura 27). En los espectrogramas de banda angosta la ventana temporal es relativamente larga, con lo que se logra una muy buena resolución en frecuencia pero una no tan buena localización de los eventos en el tiempo. Esto último es especialmente útil para la detección de formantes. En los espectrogramas de banda ancha la situación es exactamente la inversa y permiten extraer mejor parámetros como el período de entonación.

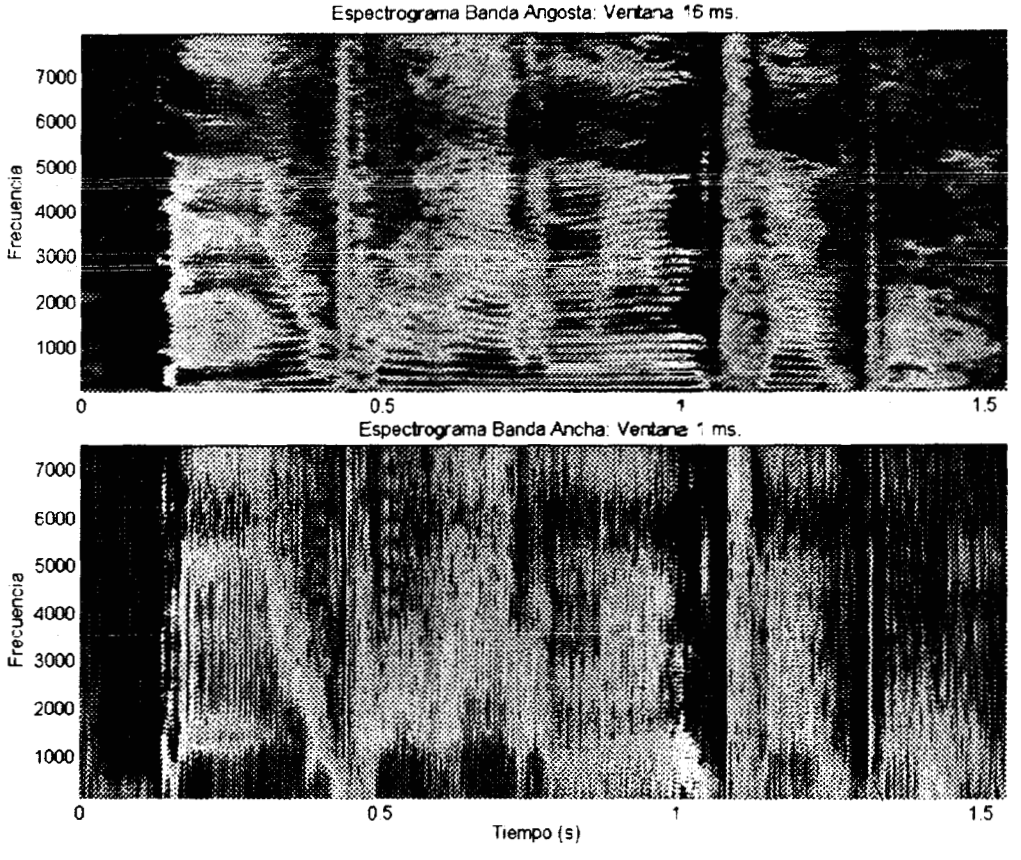

Figura 27: Espectrogramas de la frase : "The emperor had a mean temper"

Transformada Wavelet

Para evitar la limitación en resolución de la STFT se podría dejar que Δt y Δf cambien en el plano tiempo-frecuencia de manera de obtener un análisis con resolución variable (o multiresolución). Una manera de producir esto y cumplir con (5) es hacer que la resolución en el tiempo se incremente con la frecuencia central de los filtros de análisis. Más específicamente imponemos que :

$$\frac{\Delta f}{f} = c \qquad (6)$$

donde c es una constante. El banco de filtros de análisis está compuesto por filtros pasa-banda de ancho de banda relativo constante. Otra manera de ver esto es que en vez de que la respuesta en frecuencia de los filtros este regularmente espaciada en el eje de la frecuencia (como en la STFT) esta se dispone en escala logarítmica (ver Figura 28). Este tipo de bancos de filtros se utiliza, por ejemplo, para modelar la respuesta en frecuencia de la cóclea (ver capítulo sobre aspectos fisiológicos). Esto produce una resolución temporal muy buena a altas frecuencias junto con una resolución frecuencial muy buena a bajas frecuencias (Figura 29) lo que generalmente funciona muy bien para analizar las señales del mundo real (por ej. voz). En la Figura 30 se observa una confrontación entre el espectrograma y el escalograma (su equivalente en el análisis wavelets) de la sílaba 'su' (de la palabra inglesa 'suit'). Aquí la escala es inversa a la frecuencia, así que las altas frecuencias se encuentran abajo (como las componentes correspondientes a la /s/), las formantes aparecen al medio (como las bandas rojas correspondientes a la /u/), y arriba tenemos las frecuencias más bajas. Si se comparan cuidadosamente ambos análisis se pueden encontrar las correspondencias. La wavelet utilizada para generar el escalograma fue Symmlet (ver más adelante) y no se utilizó interpolación bidimensional en el gráfico (como la usada en el espectrograma) para resaltar las diferentes resoluciones tiempo-frecuencia.

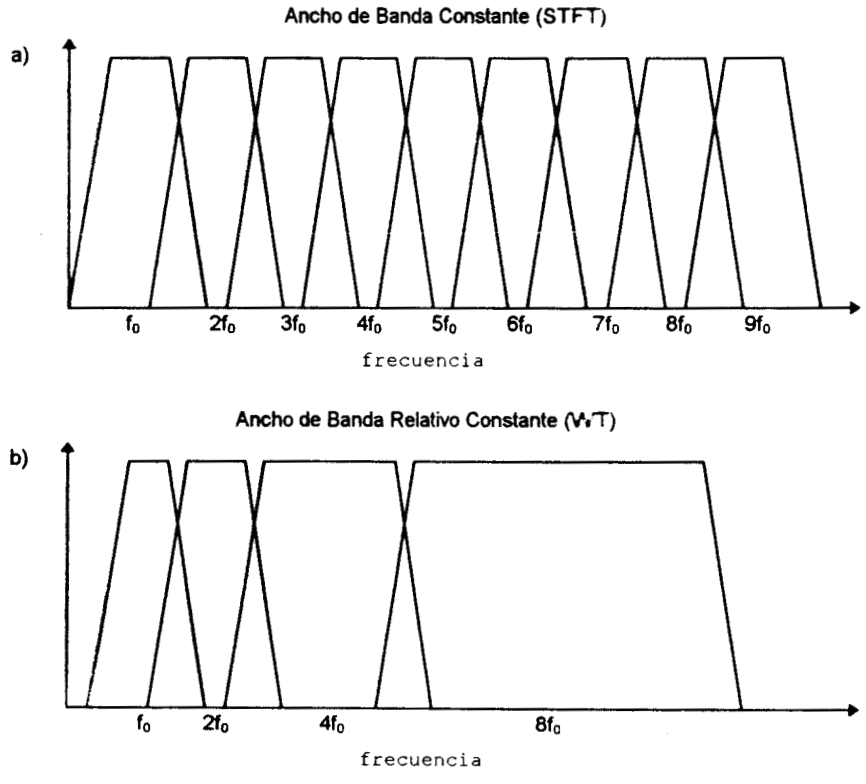

Figura 28: Distribución de los filtros en la frecuencia STFT y WT

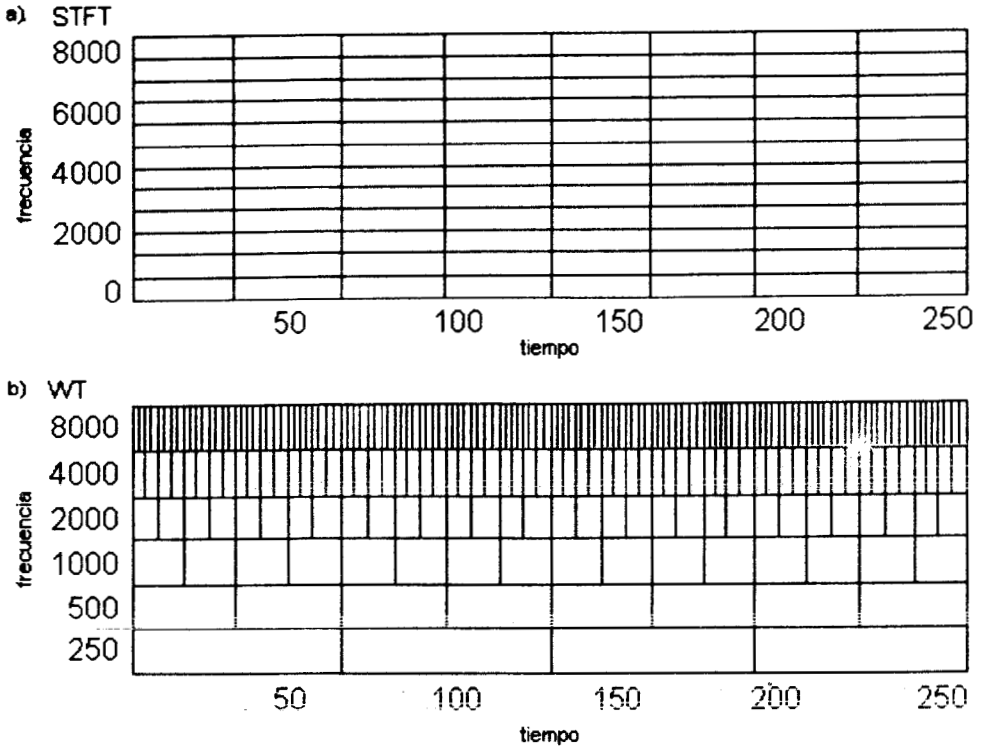

Figura 29: Resolución Tiempo-Frecuencia de la STFT y la WT

La *Transformada Wavelet Continua* (CWT) sigue las ideas anteriores agregando una simplificación : todas las respuestas al impulso de los bancos de filtros son definidas como versiones escaladas (es decirse expandidas o comprimidas) del mismo prototipo $\psi(t)$:

$$\psi_a(t) = \frac{1}{\sqrt{|a|}} \cdot \psi\left(\frac{t}{a}\right)$$

donde a es un factor de escala. Esto resulta en la definición de la CWT :

$$CWT_x(\tau, a) = \frac{1}{\sqrt{|a|}} \cdot \int x(t) \cdot \psi^*\left(\frac{t-\tau}{a}\right) \cdot dt \qquad (7)$$

Dado que se usa la misma función prototipo $\psi(t)$ (llamada wavelet básica o madre) para todos los filtros ninguna escala es privilegiada por lo que el análisis wavelet es autosimilar a todas las escalas. Además esta simplificación es útil para derivar las propiedades matemáticas de la CWT.

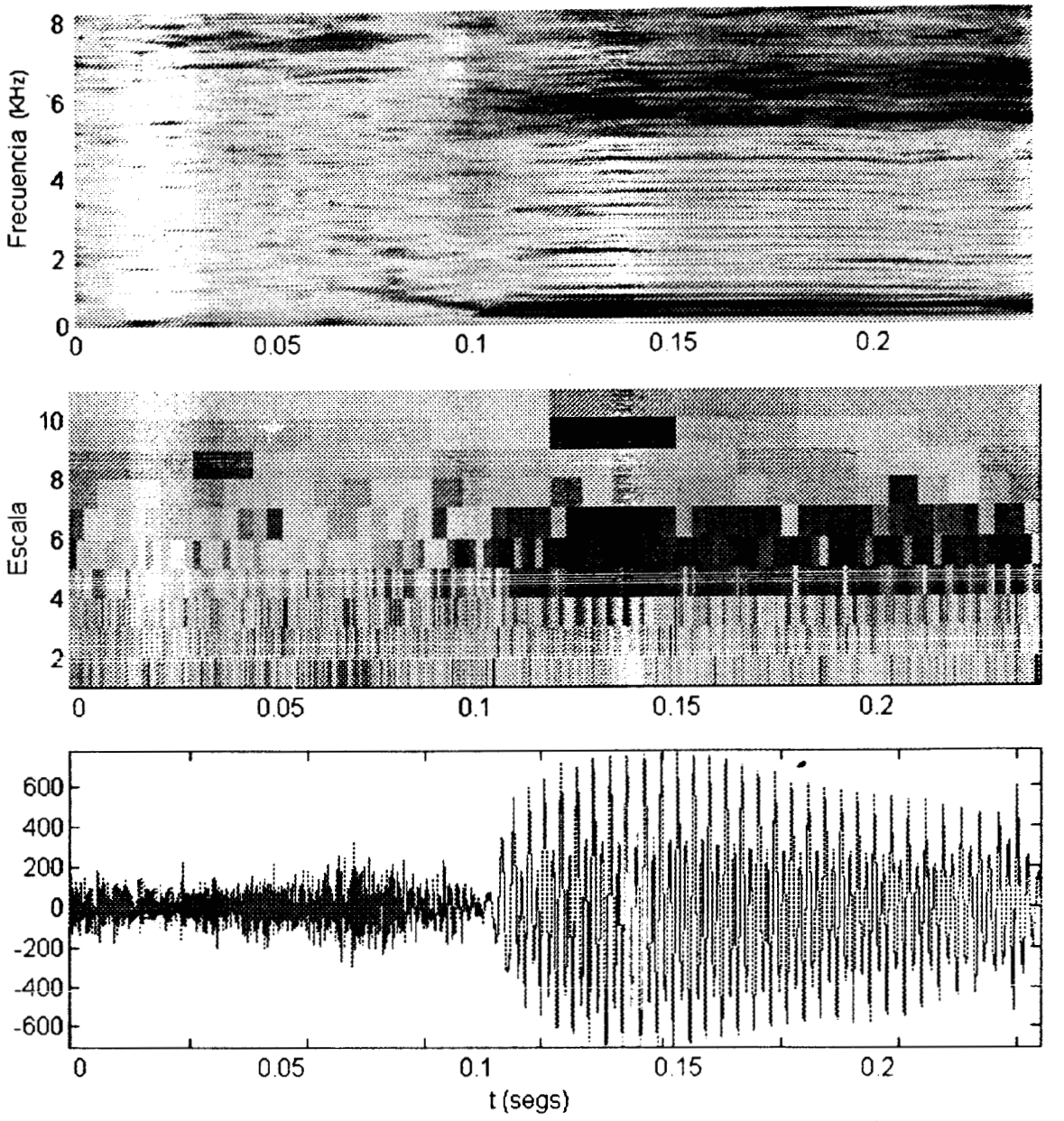

Figura 30: Espectrograma y Escalograma de la sílaba 'su' (inglés)

Para establecer relación con la ventana modulada utilizada en la STFT se puede elegir $\psi(t)$ como sigue:

$$\psi(t) = g(t) \cdot e^{-2j\pi f_0 t}$$

Entonces la respuesta en frecuencia de los filtros de análisis satisface (6) de la siguiente forma :

$$a = \frac{f_0}{f}$$

Sin embargo, de forma más general $\psi(t)$ puede ser cualquier función pasa-banda y el esquema todavía funciona. En particular uno puede evitar funciones con valores complejos y trabajar solo con aquellas que tengan valores reales.

Es importante notar aquí, que la frecuencia local $f = a \cdot f_0$, tiene poco que ver con la descripta para la STFT y ahora está asociada con el esquema de escalas. Como resultado esta frecuencia local, cuya definición depende de la wavelet madre, no está más ligada a la frecuencia de modulación sino a las distintas escalas temporales. Por esta razón se prefiere en general la utilizar el término "escala" y no "frecuencia" para la CWT. La escala para el análisis wavelet tiene el mismo significado que la escala en los mapas geográficos : grandes escalas corresponden a señales comprimidas ("vistas de lejos") mientras que escalas pequeñas corresponden a señales dilatadas ("vistas de cerca o ampliadas").

Otra manera de introducir la CWT es definir las wavelets como funciones base. De hecho, las funciones base ya aparecieron en (7) y se hacen más evidentes si la escribimos como :

$$CWT_x(\tau, a) = \int x(t) \cdot \psi^*_{a,\tau}(t) \cdot dt$$

que mide la similitud entre la señal y las funciones de la base :

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{a}} \cdot \psi\left(\frac{t-\tau}{a}\right)$$

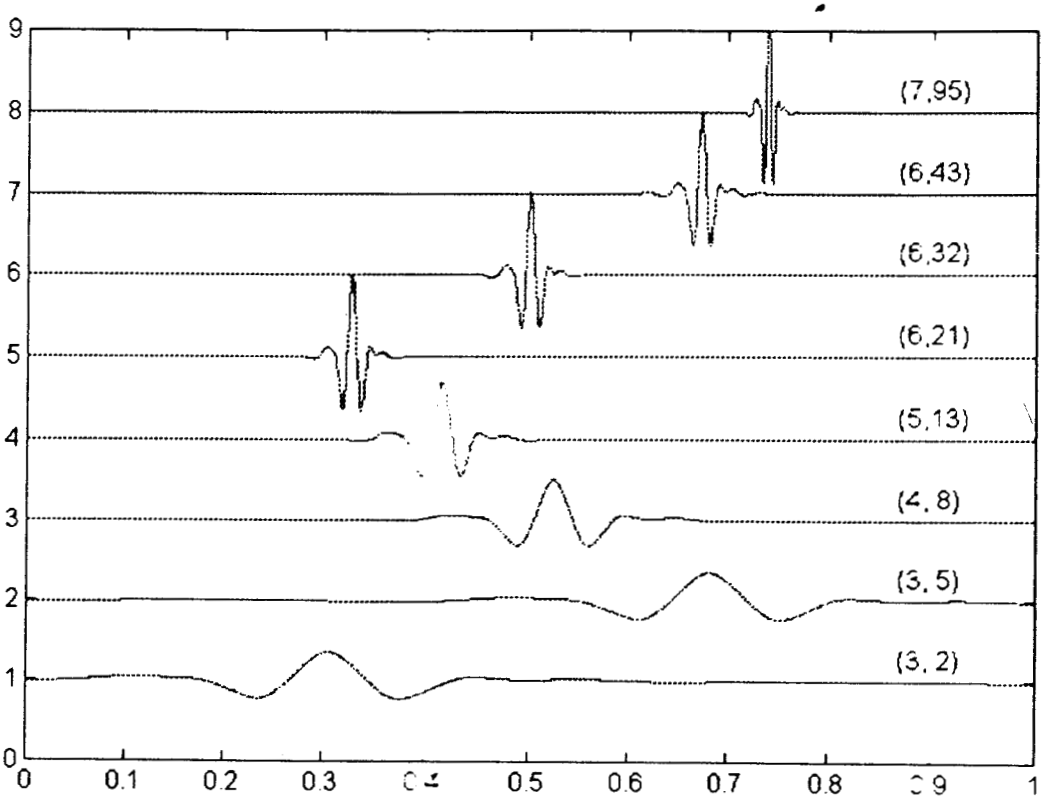

Figura 31: Wavelets Symmlets a distintas escalas y localizaciones

llamadas wavelets. Las funciones wavelets son versiones escaladas y trasladadas de la wavelet básica o prototipo $\psi(t)$. En la Figura 31 se observan algunas wavelets Symmlets a distintas escalas y localizaciones. Aquí las gráficas se realizaron de acuerdo a los parámetros (j,k), para una wavelet de ancho aproximado 2^j y localización en $k/2^j$ en el intervalo unitario.

El análisis wavelet resulta en un conjunto de coeficientes que nos indican cuán cerca está la señal de una función particular de la base. De esta manera esperaríamos que cualquier señal pudiera ser representada como una descomposición en wavelets lo que significa que $\{\psi_{a,\tau}(t)\}_{a,\tau \in R}$ debería comportarse como una base ortogonal [Mey90]. Por supuesto que este no es el caso con $\{\psi_{a,\tau}(t)\}_{a,\tau \in R}$ ya que constituyen un conjunto sumamente redundante, sin embargo aún satisfacen la fórmula de reconstrucción :

$$x(t) = c \iint_{a>0} CWT(\tau, a) \cdot \psi_{a,\tau}(t) \frac{da \cdot d\tau}{a^2} \qquad (8)$$

con la condición de que $\psi(t)$ sea de energía finita y pasa-banda. Esto implica que $\psi(t)$ oscila en el tiempo como una onda de corta duración y de ahí el nombre de wavelet u ondita. En forma más precisa si $\psi(t)$ se asume suficientemente regular (derivadas continuas hasta cierto orden) entonces la condición de reconstrucción se reduce a $\int \psi(t) \cdot dt = 0$. Esta condición es más restrictiva que la impuesta para la STFT que solo requiere que la ventana tenga energía finita.

Hasta aquí hemos visto que $\{\psi_{a,\tau}(t)\}_{a,\tau \in R}$ se comportaría para análisis y síntesis como si fuera una base ortogonal, ahora veamos si podemos conseguir una base verdaderamente ortogonal discretizando a y τ. Esto constituiría la Transformada Wavelet Discreta (DWT) y su existencia dependerá fundamentalmente de como elijamos la función $\psi(t)$ (y por supuesto de como discretizemos a y τ). En particular nos interesará la discretización de a y τ en forma diádica por su fácil implementación computacional. Con todas estas ideas presentes pasamos a definir los fundamentos teóricos del análisis multirresolución, utilizando como ejemplo el caso de la función de Haar. De allí el paso a la DWT es directo.

Fundamentos Teóricos y Definiciones

Para introducir los aspectos teóricos más importantes de nuestra herramienta utilizaremos nociones del análisis multiresolución [Str93] y comenzaremos con la función de Haar, que se puede decir que es la wavelet más simple. Asimismo, como se mencionó en la sección anterior, restringiremos nuestro análisis a la expansión en series debido a las simplificaciones matemáticas y a su aplicación computacional directa. Dada una función real $x(t)$ (o señal) en el intervalo [0,1], se puede expandir en una serie de Haar (en forma análoga a la serie de Fourier) :

$$x(t) = c_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} c_{jk} \cdot \psi(2^j \cdot t - k) \qquad (9)$$

donde $\psi(t)$ es la función definida por (Figura 32) :

$$\psi(t) = \begin{cases} 1 & si\ 0 \le t < 1/2 \\ -1 & si\ 1/2 \le t < 1 \\ 0 & de\ otra\ forma \end{cases} \qquad (10)$$

Este es un ejemplo de expansión en términos de funciones ortogonales en $L^2(0,1)$ por lo que existe una fórmula simple para los coeficientes. Pero, al contrario que la serie de Fourier, la de Haar está muy bien localizada en el espacio de manera que se puede restringir la atención a un sub-intervalo $[a, b]$ con solo tomar la suma de (9) para los índices para los cuales el intervalo $I_{jk} = [2^j k, 2^j (k+1)]$ intersecta a $[a, b]$. Además la suma parcial de la serie de Haar ($0 \le j \le N$) representa claramente una aproximación a $x(t)$ teniendo en cuenta detalles de un orden de magnitud de 2^{-N} o mayor. Estas dos propiedades : *localización en el espacio* y *escalamiento* son las más sobresalientes. Conjuntamente las funciones de Haar son creadas a partir de dilataciones y traslaciones enteras de una función ψ única (a veces llamada wavelet madre). Como ya se vio esta misma propiedad es compartida esencialmente por todas las bases wavelet y puede ser tomada como una definición aproximada de la expansión wavelet. Las expansiones wavelets que utilizaremos pueden ser vistas como generalizaciones de las series de Haar, en las cuales ψ es reemplazada por funciones más suaves.

Empezaremos con la función φ o función característica (Figura 33) en el intervalo [0, 1]. Esta es una de las funciones más simples, pero la elegimos por cumplir las siguientes propiedades :

(i) las translaciones enteras de φ, $\varphi(t - k)$, $k \in Z$, forman un conjunto ortogonal de funciones para $L^2(\boldsymbol{R})$;

(ii) φ es autosimilar. Si se divide la función en mitades, cada mitad puede ser expandida para recobrar la función original. Esta propiedad puede expresarse algebraicamente mediante la *ecuación de escala* :

$$\varphi(t) = \varphi(2t) + \varphi(2t - 1) \qquad (11)$$

Antes de decir cuales son las propiedades que les pediremos a estas funciones y como las construiremos es útil volver atrás y discutir como aparecen las funciones de Haar. El análisis será más sencillo si extendemos el dominio de nuestras funciones a todo el eje real.

Llamaremos a φ la función de escala (a veces llamada wavelet padre). El significado de la ecuación o identidad de escala es el siguiente : Sea V_0 una extensión (span) lineal de las funciones $\varphi(t - k)$, $k \in Z$. Es natural considerar este espacio teniendo en cuenta (i), dado que las funciones $\varphi(t - k)$ forman una base ortonormal para V_0. Por supuesto V_0 no abarca todo L^2, es solo el subespacio de las funciones constantes por trozos con discontinuidades de salto en Z. Podemos obtener un subespacio mayor reescalando. Sea $\frac{1}{2}Z$ el conjunto de medio-enteros $k/2$, $k \in Z$, y sea V_1 el subespacio de L^2 de las funciones constantes por trozos con discontinuidades de salto en $\frac{1}{2}Z$. Es claro que $x(t) \in V_0$ si y solo si $x(2t) \in V_1$, y

las funciones $2^{j/2}\varphi(2t - k)$ forman una base ortogonal para V_j. La ecuación de escala (11), o su versión trasladada:

$$\varphi(t-k) = \varphi(2t-2k) + \varphi(2t-2k-1)$$

nos dice que $V_0 \subseteq V_1$, dado que la base para V_0 está explícitamente representada como combinación lineal de elementos de la base de V_1.

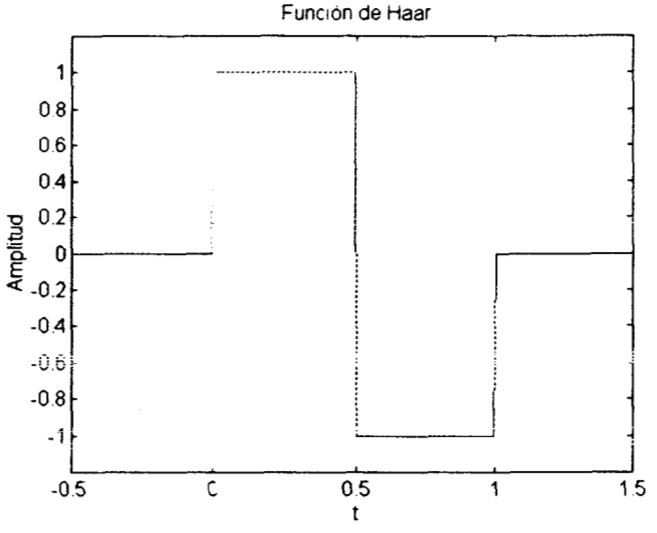

Figura 32: Función de Haar

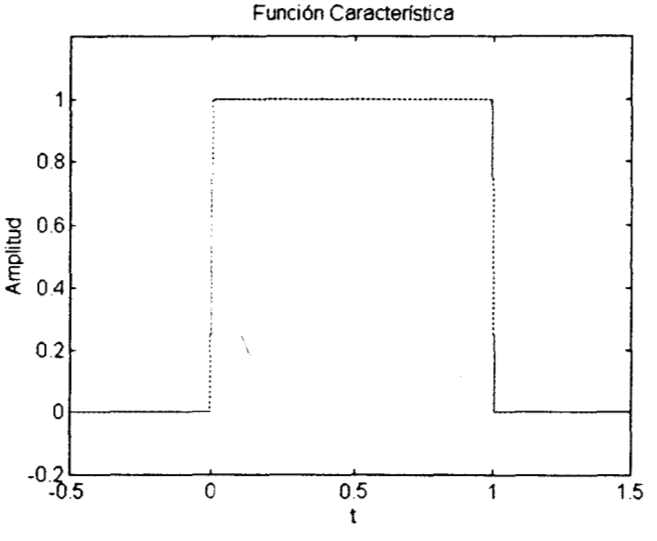

Figura 33: Función Característica

Lo anterior puede iterarse hacia arriba y hacia abajo en la escala diádica dando como resultado una secuencia creciente de subespacios V_j para $j \in Z$. Aquí V_0 consiste en las funciones de L^2 constantes por tramos con saltos en $2^j Z$, y las funciones $2^{j/2}\varphi(2^j t - k)$ para $k \in Z$ forman una base ortonormal para V_j. Podemos pasar de un espacio a otro con solo

reescalar : $x(t) \in V_j$ si y solo si $x(2^{k-j}\, t) \in V_k$ y $V_j \subseteq V_k$ si $j \leq k$. La secuencia $\{V_j\}$ es un ejemplo de lo que se denomina *análisis multiresolución*. Existen otras dos propiedades importantes de $\{V_j\}$:

$$\bigcap_{j \in Z} V_j = \{0\}, \quad (12)$$

$$\bigcup_{j \in Z} V_j \text{ es denso en } L^2 \quad (13)$$

A la vista de (13) parecería posible combinar todas las bases $\{2^{j/2}\varphi(2^j\, t - k)\}$ de V_j en una base ortonormal para $L^2(\boldsymbol{R})$. Sin embargo a pesar que $V_j \subseteq V_{j+1}$, la base ortonormal $\{2^{j/2}\varphi(2^j\, t - k)\}$ para V_j no está contenida en la base ortonormal $\{2^{(j+1)/2}\varphi(2^{j+1}\, t - k)\}$ para V_{j+1}, habiendo inclusive elementos distintos en ambas bases que no son ortogonales entre ellos.

Por lo anterior se debe encontrar otra base para $L^2(\boldsymbol{R})$. Dado que $V_0 \subseteq V_1$ y tenemos una base ortonormal para V_0 de la forma $\{\varphi(t - k)\}$ se podría tratar de completar una base ortonormal para V_1 uniendo funciones de la forma $\{\psi(t - k)\}$ para alguna función ψ. Esto equivale a preguntar por una base ortonormal de la forma deseada para el complemento ortogonal de V_0 en V_1, que denotaremos W_0, de manera que $V_1 = V_0 \oplus W_0$ (suma directa en un espacio de Hilbert).

La respuesta es simple : debemos tomar ψ de manera de ser la función generadora de Haar definida en (10). Nótese que ψ puede ser expresada en términos de φ por :

$$\psi(t) = \varphi(2t) - \varphi(2t - 1) \qquad (14)$$

lo que nos hace recordar a la identidad de escala. Ahora podemos reescalar el espacio W_0 de manera que :

$$V_{j+1} = V_j \oplus W_j \qquad (15)$$

y $\{2^{j/2}\psi(2^j\, t - k)\}_{k \in Z}$ es una base ortonormal para W_j. Si combinamos las condiciones (12), (13) y (15), obtenemos :

$$L^2(\boldsymbol{R}) = \bigoplus_{j=-\infty}^{\infty} W_j \qquad (16)$$

y dado que los espacios W_j son mutuamente ortogonales podemos juntar ahora todas las bases para W_j en una base ortogonal $\{2^{j/2}\psi(2^j\, t - k)\}_{j \in Z,\, k \in Z}$ para $L^2(\boldsymbol{R})$. Esto nos da la serie de Haar para el eje real. Utilizando las propiedades vistas podemos escribir (15) como :

$$L^2(\boldsymbol{R}) = V_0 \oplus \left(\bigoplus_{j=0}^{\infty} W_j \right) \quad (16')$$

y combinar la base $\{\varphi(t-k)\}_{k\in Z}$ para V_0 con las bases $\{2^{j/2}\psi(2^j t-k)\}_{k\in Z}$ para W_j con $j \geq 0$, para obtener una base ortonormal para $L^2(R)$. Este es el resultado al que queríamos llegar.

Propiedades del Análisis Multiresolución

Aunque los resultados anteriores se derivaron para un caso particular se pueden generalizar. De lo anterior se desprende que antes de construir las wavelets ψ debemos construir primero una función de escala φ y su análisis multiresolución asociado.

Se define un análisis multiresolución $\ldots\subseteq V_{-1}\subseteq V_0\subseteq V_1\subseteq\ldots$ con una función de escala φ como una secuencia incremental de subespacios de $L^2(R)$ que satisface las siguientes condiciones :

(i) (densidad) $\bigcup_{j\in Z} V_j$ es denso en $L^2(R)$,

(ii) (separación) $\bigcap_{j\in Z} V_j = \{0\}$,

(iii) (escalamiento) $x(t) \in V_0 \Leftrightarrow x(2^j t) \in V_j$

(iv) (ortonormalidad) $\{\varphi(t-k)\}_{k\in Z}$ es una base ortonormal para V_0.

Se sigue de la definición que $\{2^{j/2}\varphi(2^j t-k)\}_{k\in Z}$ forma una base ortonormal para V_j. Dado que $\varphi \in V_0 \subseteq V_1$ tenemos que :

$$\varphi(t) = \sum_{\gamma\in Z} a(\gamma)\cdot\varphi(2t-\gamma) \quad (17)$$

para algunos coeficientes $a(\gamma)$ que satisfacen :

$$\sum_{\gamma\in Z} |a(\gamma)|^2 = 2$$

La ecuación (17) es la análoga de (11) y por ello se denomina *identidad de escala*. Para más detalles ver [Mal89].

Transformada Wavelet Discreta

Toda la teoría anterior nos lleva a un algoritmo que permite calcular la DWT a partir de las muestras de una señal digital. El algoritmo wavelet rápido es implementado por filtrado sucesivo y submuestreo por un factor de dos (Figura 34). En cada etapa, se retienen los coeficientes wavelets correspondientes al presente nivel de resolución. El procedimiento completo es una extensión de algoritmo de Mallat para funciones base no ortogonales [Mal89]. Los filtros V(z) y W(z) son pasabajos y pasaaltos respectivamente (que corresponden a $\psi(t)$ y $\varphi(t)$), y se derivan usando las condiciones necesarias y suficientes para reconstrucción perfecta. Los de filtros generalmente se empleados son los de cuadratura

espejo (QMF) utilizados en los esquemas de filtrado sub-bandas. Esta estructura se utiliza para el análisis, para la síntesis se repite el mismo procedimiento pero hacia atrás, partiendo de los coeficientes de la transformación.

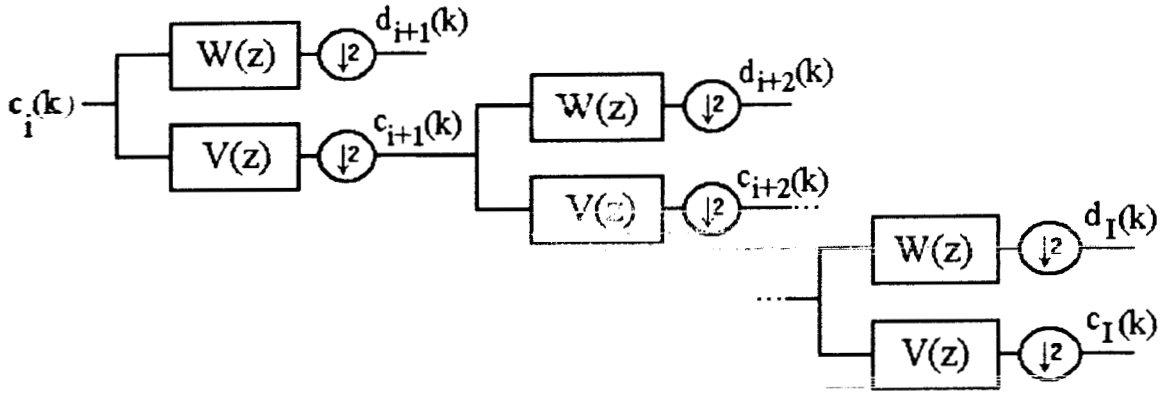

Figura 34 : Algoritmo de La DWT.

Las ecuaciones de descomposición se pueden escribir en el dominio temporal como convoluciones discretas:

$$\begin{cases} c_{i+1}(k) = \sum_l a(l-2k)\cdot c_j(l) \\ d_{i+1}(k) = \sum_l b(l-2k)\cdot c_j(l) \end{cases}$$

donde a y b representan las muestras de las respuestas al impulso de los filtros FIR (Respuesta Finita al Impulso) implicados. El proceso de submuestreo aparece porque las secuencias están muestreadas solo en los enteros pares.

Familias de Wavelets

Haar formó parte de nuestro ejemplo inicial por su simplicidad, sin embargo es discontinua y de escasas aplicaciones prácticas. Aquí describiremos brevemente las características distintivas de las principales familias de Wavelets. A veces estas características se presentan en el dominio del tiempo, aunque en varios casos las condiciones impuestas a las wavelets se llevan al dominio de la frecuencia donde algunas relaciones se simplifican.

Existen muchas posibilidades para elegir familias de wavelets que constituyan bases ortogonales. Como es imposible abarcar todas las posibilidades solo nos ocuparemos de aquellas wavelets más difundidas, y aquellas que presenten alguna característica atractiva para nuestra aplicación. Entre estas podemos citar las de Meyer, Daubechies, Symmlets,

Coiflets, Splines [Dau92] y Vaidyanathan [SoV93]. Sin embargo no existen criterios claros para escoger una familia frente a otra en una aplicación particular.

A pesar de ello se pueden notar las siguientes características que pueden ayudar a orientar nuestra búsqueda :

- Soporte Compacto : es la propiedad que asegura que la magnitud de la wavelet es igual a cero fuera de un intervalo finito.
- Simetría : en algunas aplicaciones se prefiere que exista simetría tanto en la wavelet como en la función escala.
- Regularidad : esta propiedad habla del grado de suavidad de la wavelet o la función escala.
- Localización tiempo-frecuencia : está cualidad está directamente relacionada con una mejor resolución del análisis realizado con la wavelet.

La primera propiedad tiene que ver con consideraciones prácticas que son importantes si queremos emplear el algoritmo rápido de la DWT. Para una implementación adecuada los filtros utilizados deben ser FIR y esto nos lleva a requerir que las wavelets utilizadas tengan soporte compacto. Este no es el caso, por ejemplo, con las wavelets de Meyer, que poseen soporte compacto en la frecuencia (y por consiguiente no tiene soporte compacto en el tiempo). Para realizar la DWT en este caso se debe emplear otro algoritmo totalmente distinto al presentado y con una mayor carga computacional.

Como se dijo, se prefiere que las wavelets sean simétricas, sin embargo su relación con una mejora en los resultados no esta demostrada totalmente. Si los filtros de reconstrucción y síntesis son iguales y se requiere soporte compacto entonces tal simetría es imposible (si tratamos con funciones reales). De aquí surge la idea de las Symmlets (que son las wavelets con soporte compacto menos asimétricas). Por otra parte si se deja que los filtros de reconstrucción y síntesis sean diferentes entonces las wavelets pueden ser simétricas (es el caso de las Splines Cúbicas Biortogonales).

Otra cuestión importante es lo concerniente a la regularidad (existencia de la derivada hasta cierto orden). En general se pretende que las wavelets sean lo más regulares posible (suaves) aunque tampoco existe una relación clara entre la regularidad y la "utilidad" de las wavelets.

Por último, aunque quizás lo más importante, es la localización tiempo-frecuencia de la wavelet, lo que puede observarse a través de su evolución temporal y su espectro (que deben estar bien concentrados alrededor de un punto del dominio). Una medida de la localización en frecuencia esta dada por el número de momentos desvanecientes (vanishing moments) :

$$\int_{-\infty}^{\infty} t^k \psi(t)\, dt = 0, \quad k = 0,1,\ldots,N$$

lo que es equivalente al desvanecimiento a alto orden de la transformada de Fourier en el origen :

$$\left(\frac{d}{d\xi}\right)^k \hat{\psi}_1(0) = 0, \quad k = 0,1,\ldots,N$$

Esto implica una débil forma de localización en la frecuencia debido a que la transformada de Fourier de $\psi_1(2^j t - k)$ esta mayormente concentrada alrededor de valores de $|\xi|$ del orden de 2^j.

En general la elección de los filtros que maximiza el número de momentos es distinta que la que lleva a máxima regularidad. Esto nos lleva al interrogante de cual de las dos condiciones es más importante. Otra vez esta cuestión no esta completamente clara y la respuesta depende de la aplicación específica [Dau92, cap.7].

Otras pautas pueden establecerse basadas en alguna característica especial requerida por la aplicación como ser respuesta plana en frecuencia de los filtros, filtros optimizados para algún tipo de señal, basados en modelos, etc. Recientemente han aparecido algunas familias de wavelets basadas en modelos de oído [SoC95], [SoC96], [BeT93]. Esto requiere investigación adicional acerca de los modelos en los que están basadas, para entender su relevancia en el problema planteado, por lo que se dejará como una alternativa para examinar en trabajos futuros.

En nuestras descripciones H(z) hará referencia al filtro pasabajos de análisis, RH(z) al pasabajos de síntesis, G(z) y RG(z) a los filtros pasaaltos de análisis y síntesis respectivamente. En la mayoría de los casos no existe una expresión analítica cerrada para describir las wavelets o las funciones de escala. Por esto para obtener las gráficas correspondientes se utilizó el algoritmo cascada [Dau92, cap. 6]. Este algoritmo utiliza la estructura de filtros del algoritmo rápido de la DWT con una entrada impulso para obtener una representación de las funciones implicadas con la resolución que se requiera (de acuerdo al número de etapas utilizado).

Meyer

Estas wavelets se mencionan aquí por su importancia y difusión en varias aplicaciones. Son simétricas pero, como ya se mencionó, no poseen soporte compacto lo que complica el algoritmo de calculo de la DWT. Los cálculos involucrados son completamente diferentes a los realizados en la transformada wavelet usual. Estos están basados en ventaneo, doblado, extensión y proyección en el dominio de la frecuencia, en vez de filtrado y decimación en el dominio del tiempo [Kol94].

En la Figura 35 se puede ver la forma de la wavelet de Meyer y la función de escala correspondiente, así como también sus espectros. Nótese la simetría presente en ambas y obsérvese que los espectros tienen soporte compacto (y de ahí que las wavelets no lo tengan aunque es difícil de apreciar en la figura).

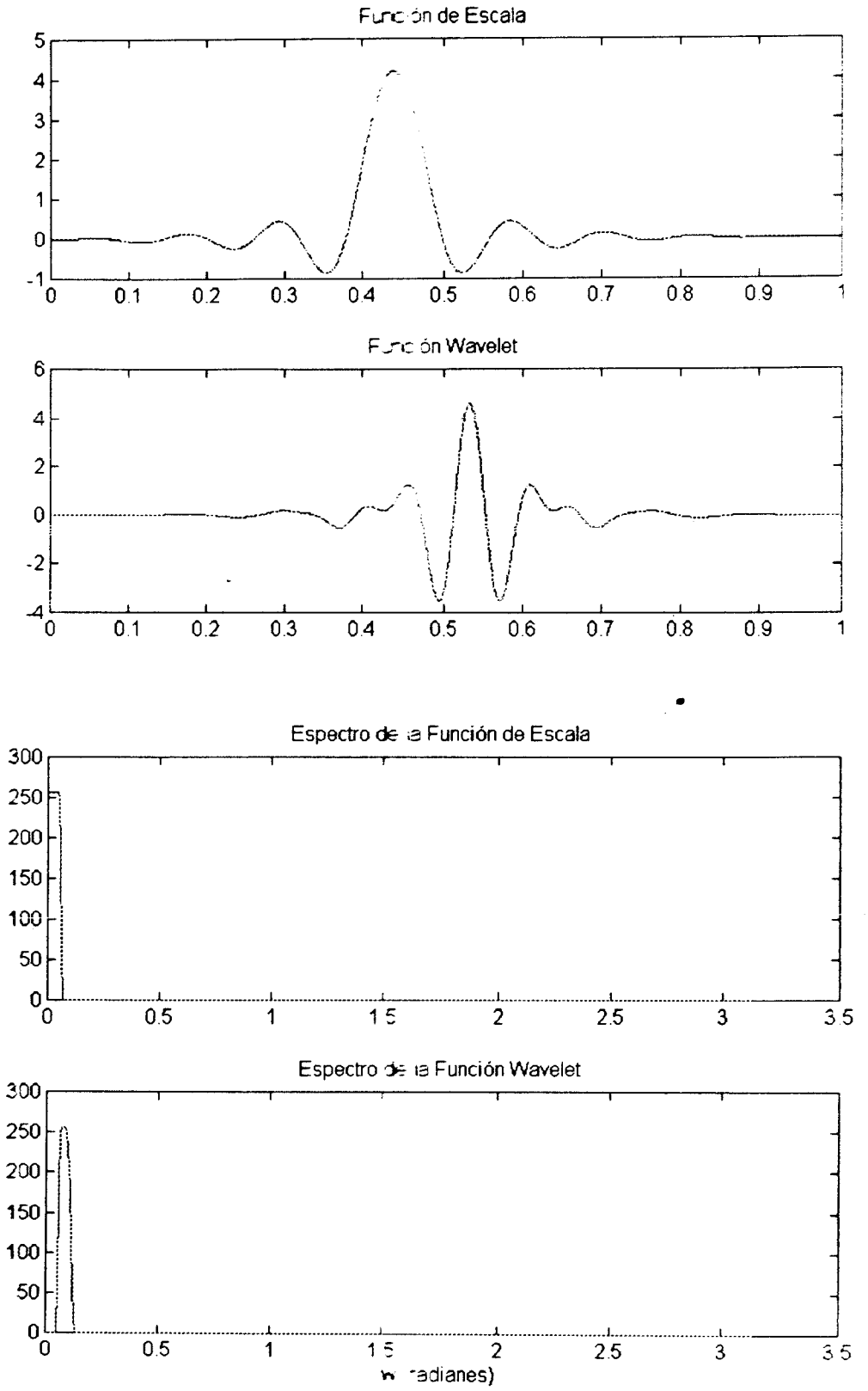

Figura 35: Wavelet de Meyer

Daubechies

Las wavelets de Daubechies maximizan la suavidad de la wavelet madre maximizando la tasa de decaimiento de su transformada de Fourier. Tienen soporte compacto y su parámetro más importante es la longitud del filtro empleado.

El algoritmo para construir esta familia de wavelets fue propuesto por Daubechies [Dau88]. Una vez seleccionado un número par de coeficientes o longitud m, se calcula el siguiente polinomio trigonométrico :

$$|Q(e^{j\omega})|^2 = \sum_{k=0}^{m/2-1} \binom{m/2-1+k}{k} \left(-\frac{1}{4}\right)^k \left(e^{-j\omega} - 2 + e^{j\omega}\right)^k$$

Para la construcción de Q a partir de $|Q|^2$, se seleccionan los ceros dentro del círculo unitario lo que conduce a una solución de fase mínima. Esto corresponde a una fuerte asimetría en ψ y φ. En este punto el filtro de síntesis pasabajos RH se puede construir a partir de los ceros dentro del círculo unitario (ya obtenidos) y los m/2 ceros en z = -1. Finalmente el resto de los filtros se calculan a partir de este utilizando las relaciones existentes entre ellos.

Cuando la longitud m es igual a 2 se reduce al caso de Haar. En la Figura 36 se observan las función wavelet y la de escala para una longitud de m=4, también se pueden ver los espectros correspondientes de los filtros generados. En la Figura 37 se puede ver lo mismo para m=20, obsérvese como las funciones se vuelven más suaves a medida que aumenta la longitud de los filtros. Nótese la asimetría presente en todos los casos.

Symmlets

El principal problema de la solución planteada en el punto anterior, para algunas aplicaciones, es la falta de simetría. Como ya se dijo no se puede tener soporte compacto y simetría si los filtros de análisis y síntesis son iguales. Una solución consiste en elegir los parámetros de los filtros de manera que las wavelets generadas sean los más simétricas posible. En este caso los filtros de cuadratura, dados también por Daubechies [Dau93], se generan igual que antes salvo que los ceros del polinomio trigonométrico son seleccionados alternativamente dentro y fuera del círculo unitario. De esta manera Symmlets son las wavelets menos asimétricas con soporte compacto y máxima cantidad de momentos. El parámetro más importante es la cantidad de momentos (m).

En la Figura 38 se puede apreciar una wavelet Symmlet con 6 momentos desvanecientes, obsérvese que a pesar de existir asimetría esta no es tan marcada como en el caso de las wavelets de Daubechies. También se pueden observar los espectros correspondientes.

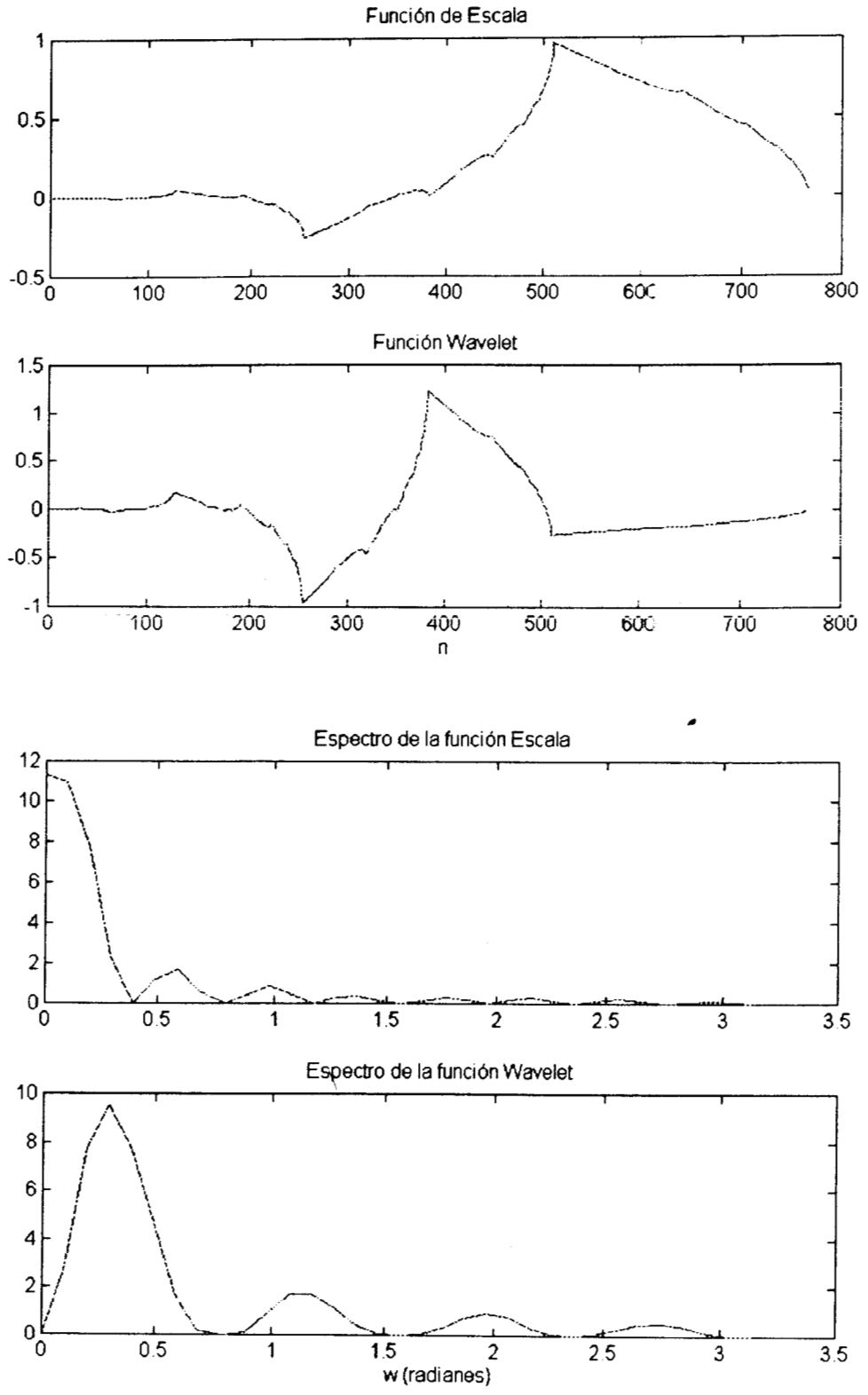

Figura 36: Wavelet Daubechies para m=4

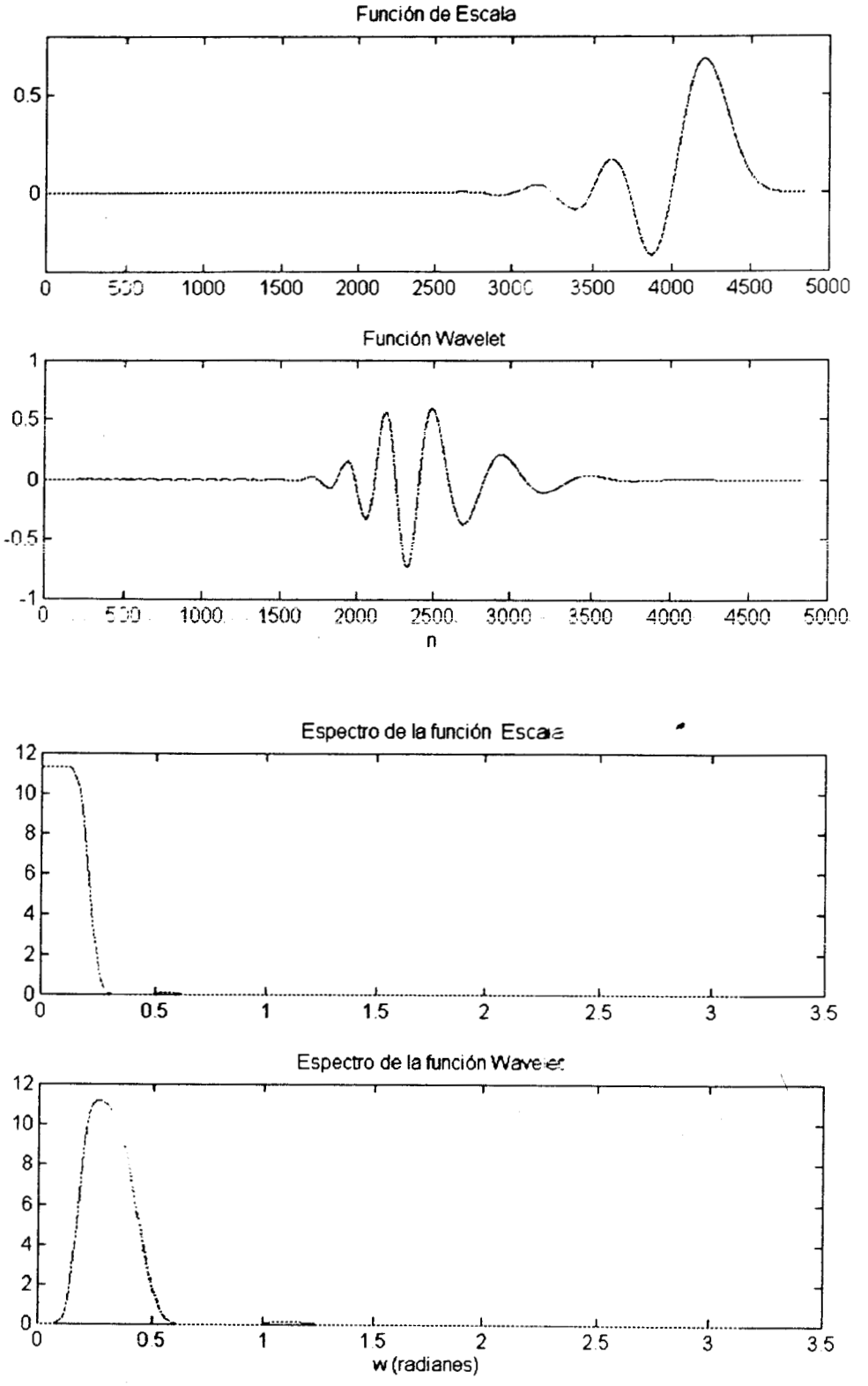

Figura 37 : Wavelet Daubechies para m=20

Coiflets

Otra de las ventajas de tener un gran número de momentos desvanecientes para ψ es que posibilita una mayor compresión debido a que los coeficientes de una función se vuelven prácticamente cero donde la función es suave. Sin embargo lo mismo no se sigue para φ. Por esta razón R. Coifman sugirió construir una base wavelet ortonormal con momentos desvanecientes no solo para ψ, sino también para φ. Estas wavelets fueron bautizadas Coiflets por Daubechies y se diseñaron para obtener m momentos desvanecientes tanto en φ como en ψ (siendo m par) [Dau92, cap. 8].

Splines

Las wavelets Splines corresponden a un esquema biortogonal basado en Splines Cúbicas de soporte compacto. Pueden alcanzar regularidad arbitraria. Tienen dos parámetros m y n que especifican el número de ceros en z = -1 requeridos para los filtros pasabajos de análisis y síntesis respectivamente, m + n debe ser par. El algoritmo utilizado es una implementación del método descripto en [Dau92] para construir wavelets splines. Primero se construye el filtro de síntesis RH a partir de m ceros en z=-1. Esto es :

$$RH(z) = \sqrt{\left(\frac{1+z^{-1}}{2}\right)^{m}}$$

A continuación H(-z) se construye a partir de n ceros en z=-1 y los ceros de P(z) :

$$H(-z) = \sqrt{\left(\frac{1+z^{-1}}{2}\right)^{m}} \cdot P(z)$$

El polinomio P(z) se calcula como en Daubechies :

$$P(z) = \sum_{k=0}^{u-1} \binom{u-1+k}{k} \left(-\frac{1}{4}\right)^{k} \left(z^{-1} - 2 + z\right)^{k}$$

donde $u = \frac{m+n}{2}$.

El filtro H(z) se calcula fácilmente invirtiendo los coeficientes de H(-z). Finalmente los filtros G y RG se calculan a partir de H y RH.

En la Figura 40 y la Figura 41 se muestran las wavelets generadas de esta manera para distintos parámetros. Aquí se puede observar como en el último caso se ha perdido la regularidad. Esto se debe a que para que las wavelets pertenezcan a C^{*} (regular hasta la derivada k-ésima) es necesario que n > 4.165 m + 5.165 (k+1). Para m y n=1 esta familia se convierte otra vez en Haar.

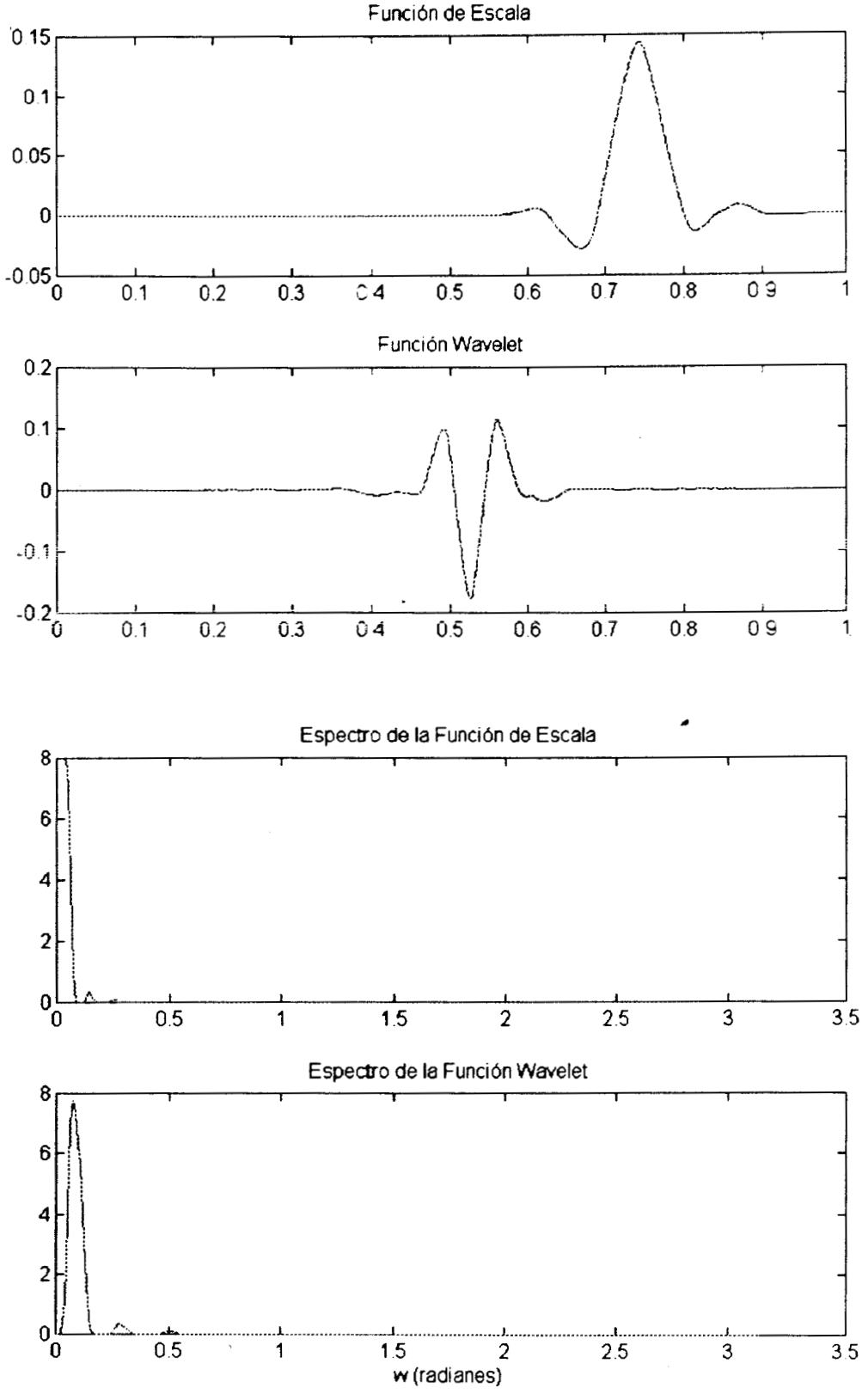

Figura 38: Wavelet Symmlet m=6

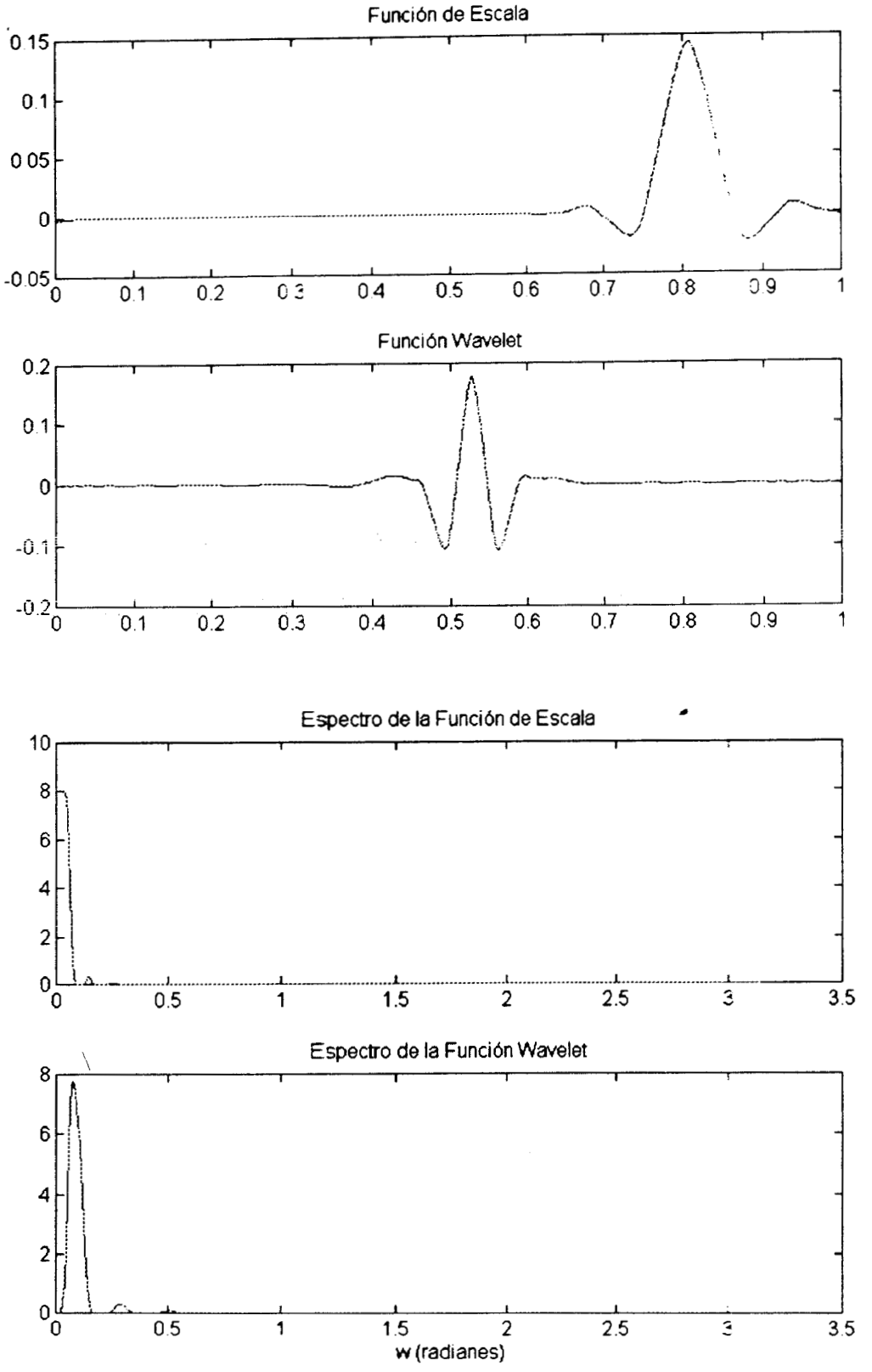

Figura 39: Wavelet Coiflet m=3

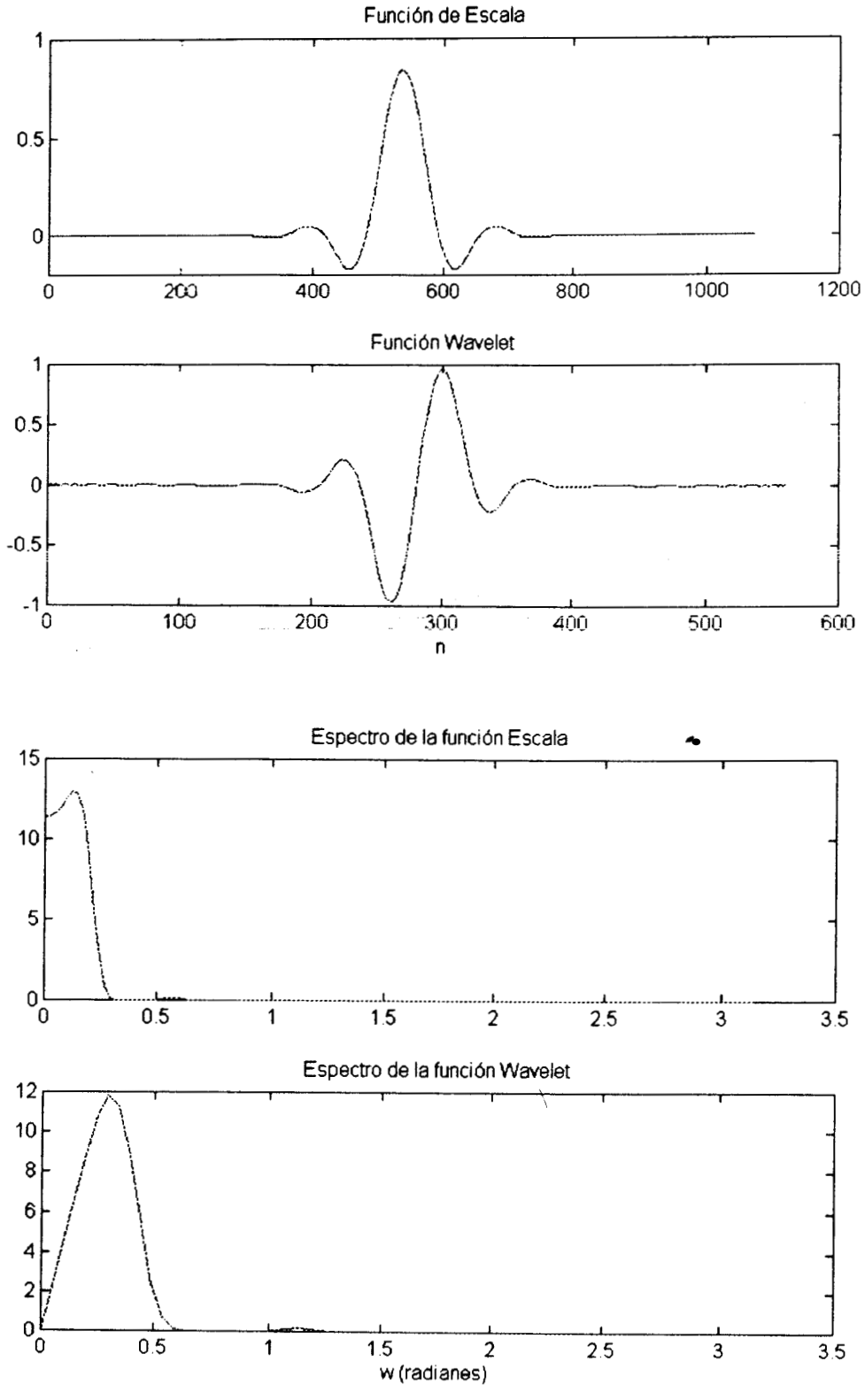

Figura 40: Wavelet Splines para m=1, n=9.

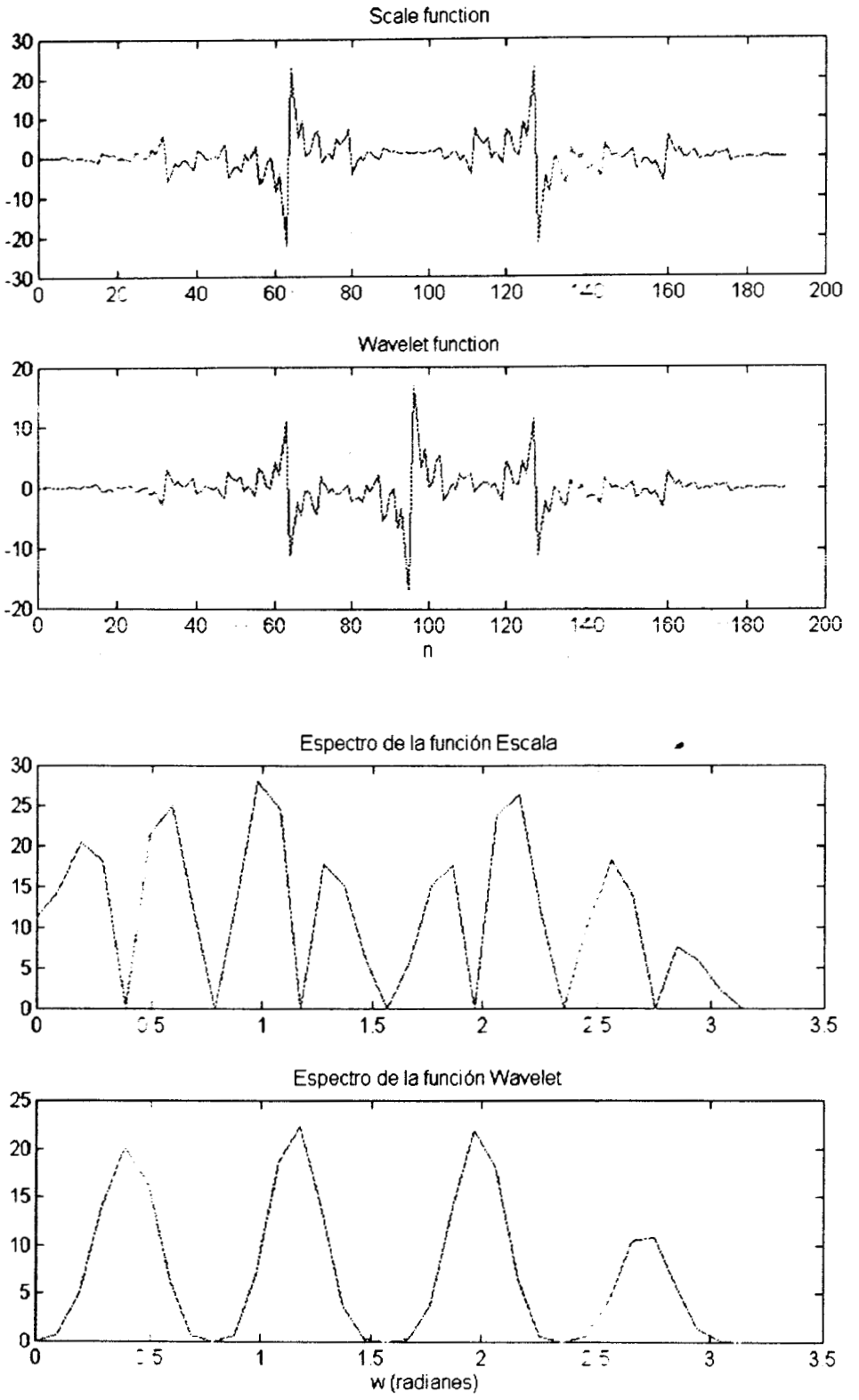

Figura 41: Wavelet Splines para m=3. n=1

Vaidyanathan

Como ya se mencionó el análisis wavelet realizado mediante la DWT está muy unido a los esquemas de filtrado sub-bandas con reconstrucción exacta utilizados en ingeniería eléctrica. El propósito de estos esquemas no es solo descomponer y volver reconstruir la señal del otro lado (más fácil y barato es utilizar un cable). El objetivo es realizar algún tipo de compresión o procesamiento entre las etapas de descomposición y reconstrucción. Para muchas aplicaciones (incluyendo voz y audio) la compresión luego del filtrado sub-bandas es más factible. La reconstrucción luego de estos esquemas de compresión ya no es perfecta, pero se espera que con filtros diseñados especialmente la distorsión sea pequeña aún para razones de compresión significativas. P.P. Vaidyanathan creó una wavelet que proporciona una reconstrucción exacta y donde los filtros han sido optimizados para codificación de voz; pero no satisface ninguna condición de momento. Para más detalles acerca de como se diseñaron los filtros ver [Vai92]

En la Figura 42 se puede observar la wavelet correspondiente y su espectro. Como se puede ver tanto la wavelet como la función escala son asimétricas.

Elección de la Base Óptima

La teoría de Wavelets u Onditas plantea un marco general para obtener familias de funciones que actúen como bases para la descomposición tiempo-frecuencia de una señal temporal [RiV91]. Ya hemos visto en la sección anterior algunas de las posibles familias o bases a utilizar. Sin embargo para cada aplicación específica, y dependiendo del tipo de señales a analizar, pueden existir bases que representen mejor a la señal (desde el punto de vista de la interpretación o la extracción de características). Es por ello que se requiere realizar una elección sobre el tipo de base a utilizar en forma objetiva. Una forma de realizar esta tarea es entrenando una red neuronal con los patrones de análisis generados por las distintas bases en evaluación. Existen trabajos recientes donde se han realizado comparaciones de este tipo entre Wavelets y Fourier [Fav94-3]. Sin embargo existe la posibilidad de utilizar criterios más directos para la elección de un base. Una posibilidad consiste en utilizar técnicas como Wavelets Packets o Matching Pursuit.

En el caso de Wavelets Packets [Wic91b], [Cod94] se plantea un marco más general para la descomposición de una señal en términos de una base ortogonal (donde la base wavelets es un caso particular). Luego la base elegida depende de un costo asociado al análisis o descomposición logrado por cada base posible. Sin embargo las bases son generadas a partir de una función wavelet única por lo que volvemos al problema inicial. Este enfoque ha sido muy útil en aplicaciones de compresión y filtrado de señales [Tas95] pero requeriría una adaptación para el caso de clasificación [SaC94].

En Matching Pursuit [MaZ93] se utiliza un diccionario que puede estar formado por varias bases. La señal a analizar se compara a través de la correlación con los elementos del diccionario. El elemento más parecido se resta de la señal y el proceso continua hasta que el residuo es suficientemente pequeño. En este caso no se está atado a una base determinada aunque el proceso puede ser muy lento para diccionarios grandes.

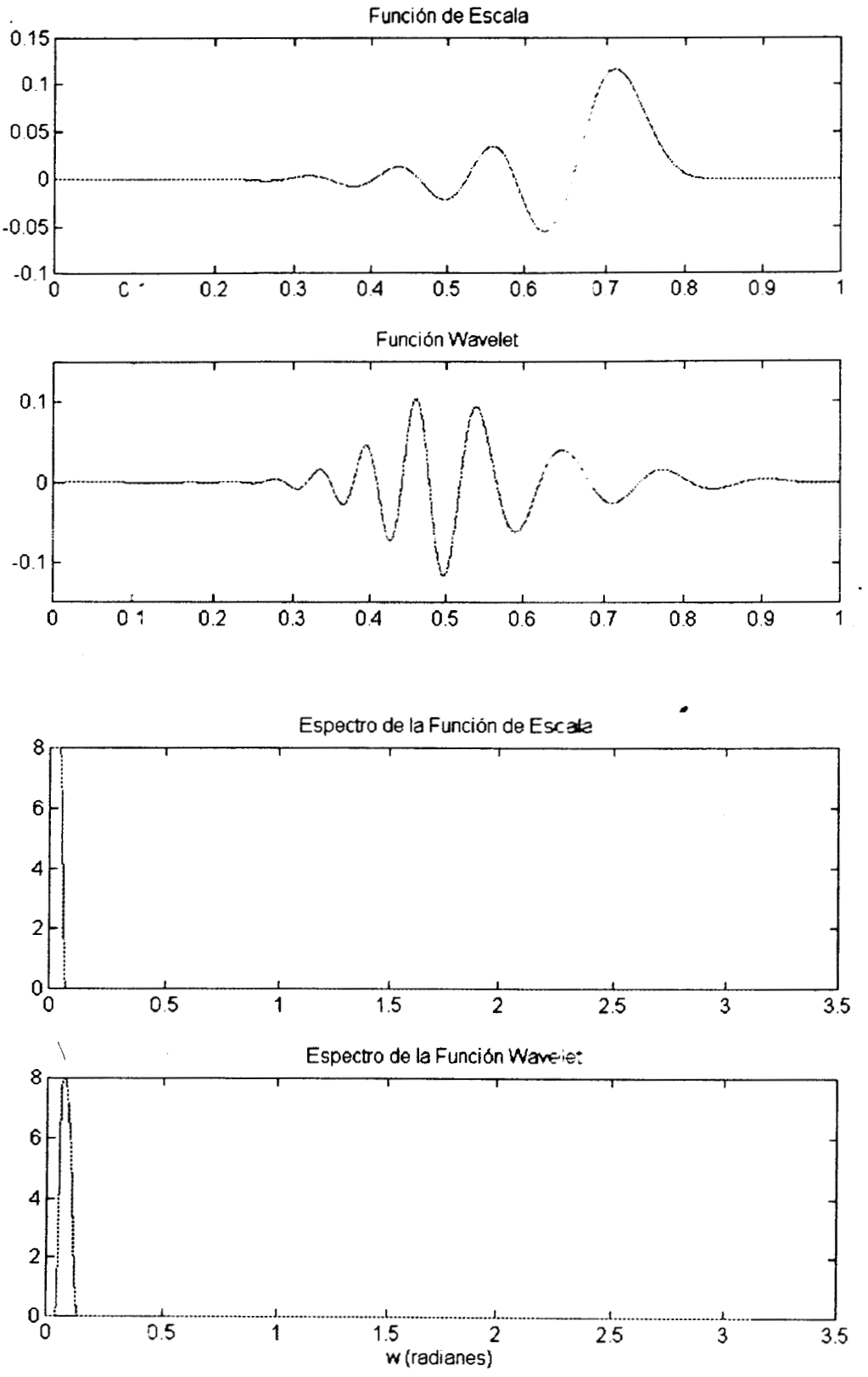

Figura 42: Wavelet de Vaidyanathan

Otra posibilidad sería tratar de describir el espacio de clases generado por cada base mediante métodos basados en agrupamiento, de manera de poder medir de alguna manera la separación entre las clases. Esta posibilidad se explora en el capítulo siguiente.

Aspectos de Implementación Práctica

En esta sección describiremos los aspectos relacionados con la implementación de los algoritmos y los parámetros utilizados en los experimentos.

En el caso de Fourier las pruebas se realizaron utilizando una ventana de Hamming de 256 puntos que se dezplazó en pasos de 128 puntos, generando 128 coeficientes espectrales cada 8 mseg (como ya se indicó las señales fueron muestreadas a 16 KHz). Estos coeficientes se llevaron a escala logarítmica como es usual en los espectrogramas clásicos para resaltar la energía de las bandas de alta frecuencia.

Otro experimento se llevó a cabo reduciendo estos 128 coeficientes a solo 20 aplicando escala de Mel a los mismos. Esta escala es logarítmica en la frecuencia y se basa en estudios psicoacústicos de la percepción de la frecuencia. En la Figura 43 se pude observar la relación con la escala lineal de frecuencias en el rango considerado. En la Figura 44se puede apreciar el proceso completo para el fonema /jh/ en sal.

En el caso de wavelets se tomó una ventana rectangular de 128 coeficientes u 8 mseg para poder comparar con el caso de Fourier en condiciones similares esto produjo un total de 7 escalas que iban desde la banda 4-8 KHz (con 64 coeficientes) hasta la banda 125-62.5 Hz (con un solo coeficiente). Lo mismo se repitió para cada familia de wavelets utilizada : Meyer, Daubechies, Splines, Haar y Vaidyanathan. Otra posibilidad ensayada consiste en agrupar los coeficientes en níveles o bandas de frecuencia y pasarlos al clasificador a distinta velocidad para cada nivel (ya existen muchos más coeficientes por unidad de tiempo en las bandas de frecuencia más alta). A esta última opción se la ha denominado en otros trabajos agrupamiento en frecuencia y a la primera agrupamiento temporal [FaG94]. Aquí también se utilizó escala logarítmica para la magnitud de los coeficientes, ya que en experimentos preliminares esto produjo mejores resultados que la escala lineal.

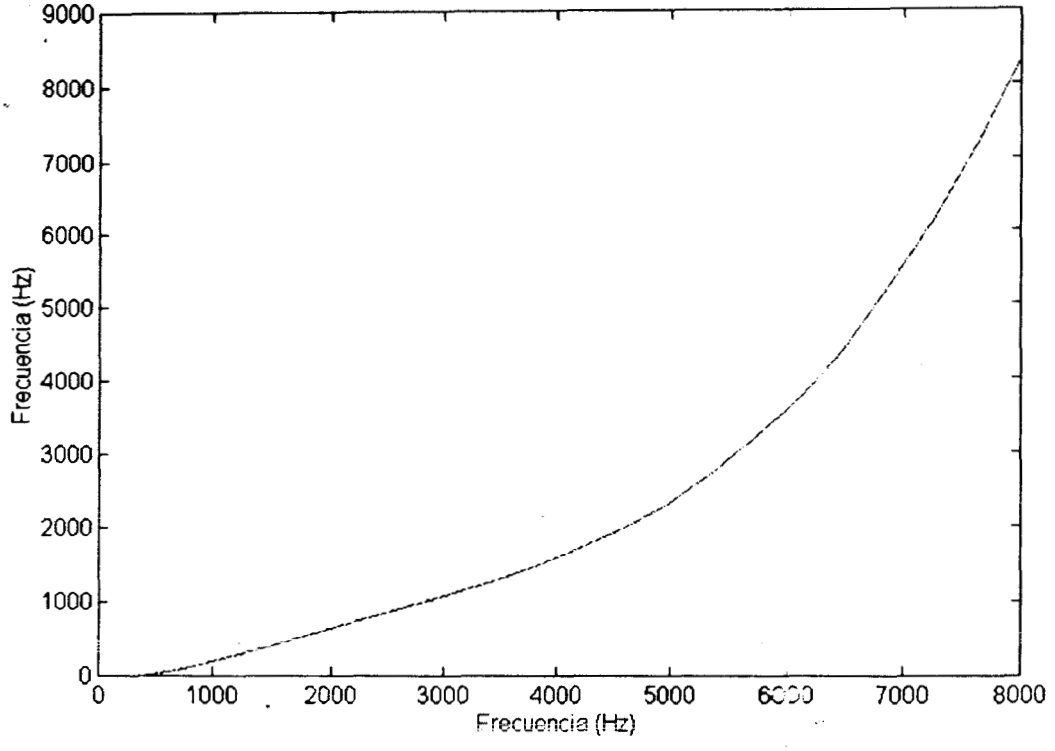

Figura 43: Escala de Mel

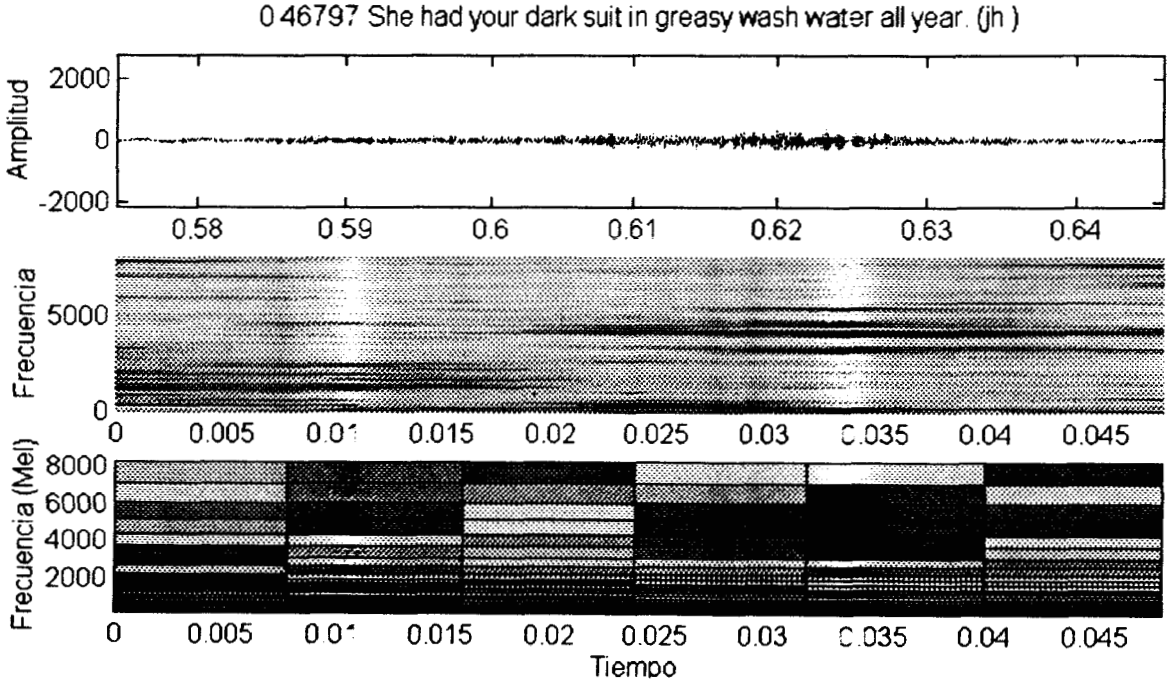

Figura 44: Conversión del Espectrograma a Escala de Mel

V . El Clasificador

Introducción

Como se mencionó en la introducción general de la tesis una de las etapas de un sistema de RAH lo constituye el clasificador de fonemas. Aunque nuestro trabajo no pretende optimizar esta etapa, la vamos a utilizar para 'medir' y comparar la eficacia del análisis realizado con Fourier y Wavelets. Además, como se mencionó en la sección relativa al procesamiento, utilizaremos una técnica basada en agrupamiento como parte de un criterio para elegir la wavelet más adecuada para el análisis del habla. Por ello en esta sección describiremos los distintos métodos existentes para realizar esta tarea y la razón para escoger el o los métodos propuestos en nuestro trabajo.

El problema de clasificación o reconocimiento de patrones tiene una larga historia. Las técnicas que han dominado el reconocimiento de patrones durante mucho tiempo están basadas en la estadística [DuH73]. Sin embargo, en los últimos años se han intentado aplicar otras que han surgido de la *Inteligencia Artificial* (IA) y constituye un área denominada *Aprendizaje Maquinal* (AM).

Dentro de la amplia variedad de técnicas utilizadas en AM debemos diferenciar entre las que utilizan *Aprendizaje Supervisado* (AS) y las que utilizan *Aprendizaje No Supervisado* (ANS). En el caso del AS se emplean un conjunto de ejemplos en la forma de una serie de atributos o características y una etiqueta que define la clase a la que corresponde cada ejemplo. En el caso del ANS los ejemplos no necesitan estar 'etiquetados' ya que el método se encarga de agrupar a los ejemplos o patrones en las clases o grupos 'naturales'.

Entre las técnicas de AS encontramos el paradigma de aprendizaje por reglas inducidas a partir de un conjunto de ejemplos de entrenamiento. En este paradigma el algoritmo de aprendizaje busca una colección de reglas que clasifican "*mejor*" los ejemplos de entrenamiento. El término *mejor* se puede definir en términos de exactitud y comprensión de las reglas generadas. Dichas reglas con frecuencia pueden ser obtenidas a partir de un árbol de decisión (AD). Este puede pensarse como un diagrama de flujo en donde cada nodo representa una prueba y cada rama que sale del nodo representa un resultado posible de dicha prueba. Dentro de estas técnicas se encuentran ID3, C4.5 y CART entre las más utilizadas.

Aunque los AD son intuitivamente atractivos y han tenido aplicaciones exitosas [Qui93], [BFO84] existen algunos problemas que pueden obstaculizar su empleo en casos reales. Problemas tales como el diseño del árbol en si mismo, la presencia de datos inconclusos, incompletos o ruidosos y el empleo simultáneo de todos los vectores de atributos. Más aún, existen algunas limitaciones debidas al hecho de que solo pueden describir el espacio de clases en términos de hiperplanos perpendiculares a los ejes de dicho espacio. Esto provoca que para describir regiones complejas se requiere un número muy grande de reglas o nodos

en los árboles. Otro problema, para nuestro caso, es que existe muy poco trabajo desarrollado acerca de la aplicación de estas técnicas a la clasificación de patrones dinámicos o dependientes del tiempo.

Una manera diferente de atacar el problema de reconocimiento de patrones es con el uso de *Redes Neuronales Artificiales* (RNAs), siendo la alternativa más común los *Perceptrones Multicapas* (PMC). Estos tienen la ventaja de que pueden aprender fronteras de decisión arbitrariamente complejas y se entrenan por métodos de AS [Lip87].

Las RNAs están basadas parcialmente en la estructura y funciones del cerebro y sistema nervioso de los seres vivos. Una RNA es un sistema de procesamiento de información o señales compuesto por un gran número de elementos simples de procesamiento, llamados *neuronas artificiales* o simplemente *nodos*. Dichos nodos están interconectados por uniones directas llamadas *conexiones* y cooperan para realizar procesamiento en paralelo con el objetivo de resolver una tarea computacional determinada.

Una de las características atractivas de las RNAs es su capacidad para auto-adaptarse a condiciones ambientales especiales cambiando sus fuerzas de conexión llamadas *pesos*, o su estructura.

Muchos modelos de RNAs han sido desarrollados para una variedad de propósitos. Cada uno de ellos difieren en estructura, implementación y principio de operación; pero a su vez tienen características comunes. Las RNAs Anteroalimentadas (llamadas así porque las conexiones son siempre hacia adelante) no poseen realimentaciones y están principalmente orientadas a la clasificación de patrones estáticos. Sin embargo es posible extender varios conceptos de este tipo de redes a las RNAs Recurrentes que permiten distintos grados de realimentación en su estructura.

También existen RNAs que utilizan métodos de ANS para su entrenamiento. Entre estas se encuentran las redes de Kohonen [Lip87] que utilizaremos en nuestro problema acerca de la elección de la mejor familia de wavelets para la tarea de clasificación.

Otra alternativa es la utilización de técnicas híbridas que aprovechen las ventajas de cada método [Set90], [Set91], [Bre91]. Así surge la idea de los árboles de redes neuronales, en los cuales cada nodo de un árbol está formado por una red neuronal [SaM91]. También existe la posibilidad de mejorar el diseño de una red utilizando algunas relaciones existentes entre árboles y redes [GMM95].

Por todo lo expuesto en nuestro trabajo se eligió trabajar con RNAs, en particular con aquellas arquitecturas y algoritmos que permitan tratar los aspectos dinámicos de la señal de voz.

Este capítulo se organizará de la siguiente forma. Primero se describirá brevemente el PMC y el algoritmo de Retropropagación, por su importancia en la clasificación de patrones estáticos. A continuación se presentarán las arquitecturas para clasificación de patrones dinámicos y los algoritmos necesarios para su entrenamiento, derivados en su mayoría del algoritmo de entrenamiento de los PMC. Luego se presentarán las redes de Kohonen y se

explicará el criterio empleado en la selección de las wavelets. Finalmente se describirán los criterios seguidos para la elección de las arquitecturas o algoritmos utilizados en el trabajo y los aspectos prácticos de su implementación.

Redes Neuronales Estáticas : Perceptron Multicapa

El PMC consiste en un arreglo de nodos ubicados en capas, de forma tal que los nodos de una capa están conectados a todos los nodos de la capa anterior y de la siguiente mediante los pesos de conexión. La primer capa se denomina *capa de entrada* y la última se denomina *capa de salida*, las capas que quedan entre estas 2 se denominan *capas ocultas* (Figura 45). No existe un límite para fijar la cantidad de capas de un PMC, pero se ha demostrado que un PMC con una capa oculta y con un número suficiente de nodos es capaz de solucionar casi cualquier problema. Si se agrega una capa oculta más, un PMC soluciona cualquier tipo de problemas y en forma más eficiente que con una sola capa oculta. El decir que puede solucionar cualquier problema se refiere a la capacidad de estas redes de aproximar regiones de decisión de complejidad arbitraria. En la Figura 46 se observa un esquema simplificado del PMC mostrado en la Figura 45, este tipo de esquemas serán los que utilizaremos en la descripción de las distintas arquitecturas.

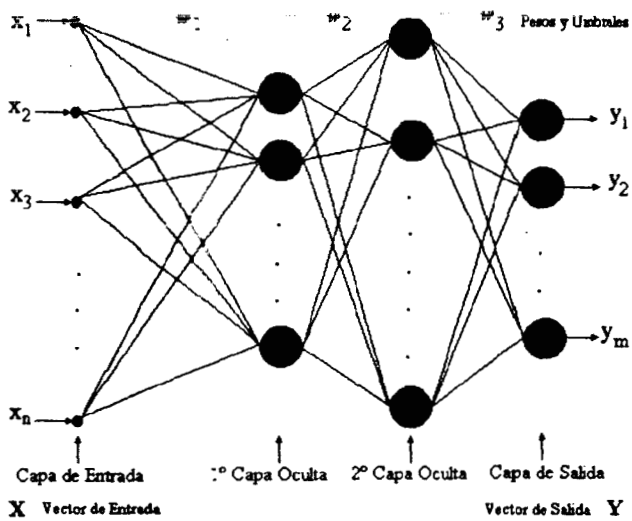

Figura 45: PMC con dos capas ocultas.

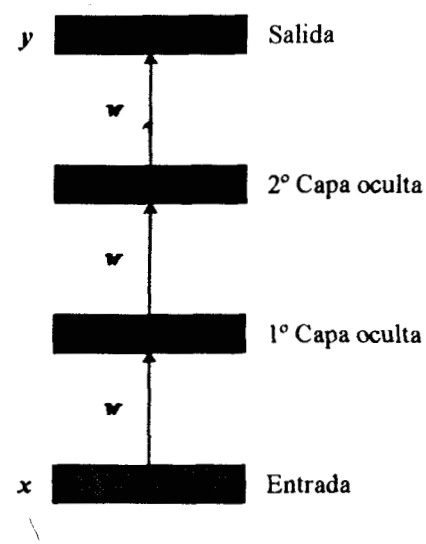

Figura 46: Esquema Simplificado

En la capa de entrada, los valores pasan directamente a la siguiente capa sin cambiarlos (por eso en los diagramas los nodos correspondientes a las entradas aparecen distintos). Para entrenar el PMC generalmente se utiliza el algoritmo de retropropagación [Lip87]. Durante el entrenamiento, se presenta el valor de las variables de entrada para cada, y la respuesta de cada nodo en la capa de salida es comparada con la correspondiente respuesta deseada. Los errores asociados son calculados para adaptar los pesos en cada capa por medio de un método de gradiente descendente. Como consecuencia de este método, la función de activación de los nodos debe ser diferenciable. Dichas funciones de activación tienen la siguiente forma: $y(x) = \frac{1}{1-e^{-x}}$, donde x es la sumatoria de las entradas al nodo pesadas por

los pesos de conexión y *y(x)* la salida del nodo correspondiente. Este proceso es repetido hasta que el error entre los valores de salida y los deseados, para cada uno de los casos, llega a un nivel aceptable. En síntesis, los pasos implicados en el aprendizaje serían los siguientes:

```
procedimiento retropropagación
{
inicializar pesos y umbrales
mientras (criterio de terminación no se alcance)
{
            seleccionar ejemplo
            aplicar a la entrada de la red
            calcular salida de la red
            calcular error entre la salida y su valor deseado
            propagar el error hacia atrás
            ajustar pesos y umbrales para minimizar el error
}
}
```

La fórmula de actualización de los pesos de la última capa tiene la siguiente forma :

$$w_{i,j}(n) = w_{i,j}(n-1) + \eta \cdot error_j(n) \cdot y'_j(n) \cdot x_i(n)$$

Donde $w_{i,j}$ es el peso que va del nodo i al nodo j, η es el coeficiente de aprendizaje, x_i es la entrada correspondiente, y'_j es la derivada de la función de activación evaluada en Σx_I y n es el instante de tiempo considerado. En el caso que se trate de una capa distinta de la última entonces el error se estima propagando hacia atrás los errores cometidos en la capa de salida :

$$error'_i(n) = \sum_j error_j(n) \cdot w_{i,j} \cdot y'_j(n)$$

La idea fundamental detrás del método de gradiente descendente puede verse a través de una analogía con una bola que se desliza por una superficie. La altura de la superficie representa el error para cada punto en el espacio de los pesos. De esta manera la bola se desliza por la superficie siguiendo el sentido del gradiente descendente de la misma. Como se puede apreciar de la analogía, la posición inicial de la bola es de suma importancia. Si estamos en una zona relativamente plana de la superficie tardaremos mucho en hallar un mínimo. Si estamos cerca de un mínimo local, es muy probable que caigamos en él. Para solucionar parcialmente este problema se agrega al algoritmo un término de inercia o momento. También se puede repetir varias veces el procedimiento iniciándolo en distintos puntos (generalmente cerca del origen) y guardar los pesos de la corrida que llegue a menor error final. Para más detalles se puede consultar la extensa bibliografía al respecto (por ej. [Lip87], [WiL90], [Has95]).

Mientras que el PMC presenta las ventajas anteriormente mencionadas, también presenta el inconveniente que debe definirse su estructura en términos de nodos y capas, y en general puede demorar demasiado para llegar a un mínimo aceptable. Otro inconveniente de esta técnica al aplicarla al caso del habla es que está orientada a la clasificación de patrones estacionarios.

Redes Neuronales Dinámicas

La aparición de técnicas de entrenamiento eficaces para redes neuronales -en particular las redes anteroalimentadas- permitió la aplicación de las mismas al procesamiento del habla, aunque hasta hace muy poco tiempo estuvieron orientadas a patrones estacionarios. Para evitar este escollo se diseñaron redes neuronales que - además de los patrones estáticos - incorporaran simultáneamente información que fuera generada en diferentes instantes. La utilización de redes con retardos (TDNN) permite descubrir características acústico-fonéticas y sus relaciones a lo largo del tiempo sin verse afectadas por las traslaciones del patrón de entrada [WHH89]. Una red neuronal que no es invariante a la traslación requiere una segmentación precisa para alinear el patrón (frame) de entrada en forma adecuada. Esta invariancia al desplazamiento ha sido considerada un factor determinante en algunos modelos neuronales que han sido propuestos para reconocimiento de secuencias fonéticas y es por lo tanto fundamental para nuestra aplicación. Otra arquitectura con características similares para la clasificación de patrones dinámicos son las redes neuronales recurrentes (RNN). Existen varias simplificaciones que se pueden aplicar para extender el algoritmo de retropropagación para entrenar estas redes.

Extensión de Retropropagación para Aprendizaje Temporal

Como se dijo el enfoque clásico y la mayoría de las aplicaciones de redes neuronales están orientadas al tratamiento y clasificación de patrones estáticos [Hau95]. En estos casos la red adquiere un mapeo estático para producir un patrón espacial de salida como respuesta a un determinado patrón espacial de entrada. Sin embargo en muchas aplicaciones (de ingeniería, científicas, económicas, etc.) se requiere modelar un proceso dinámico donde se requiere una secuencia temporal en respuesta a determinada señal temporal de entrada. Este es el caso que nos ocupa en el problema de reconocimiento del habla, donde en base a una señal (que pueden estar dada por los espectros o la transformada Wavelet) se genera la secuencia probable de fonemas que emitió el hablante. Otro ejemplo es el modelado de una planta para aplicaciones de control. El modelo resultante en ambos casos se denomina *red de asociación temporal*.

Las redes de asociación temporal deben poseer una arquitectura recurrente para poder manejar la naturaleza dependiente del tiempo de las asociaciones. Por ello sería muy útil extender las redes anteroalimentadas multicapas y sus algoritmos de entrenamiento asociados (como retropropagación) al dominio temporal. En general esto requiere redes con conexiones de realimentación y mecanismos apropiados de aprendizaje.

Dos casos especiales de las redes de asociación temporal son las de reproducción de secuencias y las de reconocimiento de secuencias. Para la reproducción de secuencias , una red debe ser capaz de generar a partir de un trozo de una secuencia, el resto de la misma. Esto es apropiado para predecir tendencias (por ejemplo en el caso económico) a partir de la historia pasada , o predecir el curso futuro de una serie de ejemplos en base a los ejemplos. En el caso de reconocimiento de secuencias, una red debe producir un patrón espacial o una salida fija en respuesta a una secuencia de entrada específica. Este es el caso del reconocimiento del habla.

En lo que sigue trataremos arquitecturas de redes neuronales con distintos grados de recurrencia y los métodos de aprendizaje correspondientes capaces de procesar secuencias temporales.

Redes Neuronales con retardos temporales

Las Redes Neuronales con retardos (TDNN) consisten en unidades elementales similares a las de un perceptron, pero modificadas a fin de que puedan procesar información generada en distintos instantes. A las entradas sin retardos de cada neurona, se les agregan las entradas correspondientes a instantes distintos. De este modo, una unidad neuronal de estas características es capaz de relacionar y procesar conjuntamente la entrada actual con eventos anteriores.

Considere la arquitectura mostrada en la Figura 47. Esta red mapea una secuencia finita temporal $\{x(t), x(t-\Delta), x(t-2\Delta),..., x(t-m\Delta)\}$ en una salida única y (se puede generalizar para el caso en x y y sean vectores (Figura 48)). Se puede ver esta red como un filtro de tiempo discreto no lineal (FIR).

Esta arquitectura es equivalente a una red neuronal anteroalimentada con una capa oculta que recibe un patrón espacial con $(m+1)$ dimensiones x generado por un línea de retardo a partir de la secuencia temporal. Esto es, si los valores deseados para la salida son especificados para distintos tiempos t, entonces se puede utilizar el algoritmo de retropropagación para entrenar esta red como un reconocedor de secuencias.

Este tipo de redes ha sido aplicado exitosamente al reconocimiento de voz [TaH87], [ElZ88], [WHH89], [Lip89] . En las TDNN de Waibel se incluye generalmente una capa oculta extra (también con retardos) y se entrenan por el algoritmo de retropropagación a través del tiempo. De esta forma cada unidad neuronal con retardos posee la capacidad de extraer relaciones temporales entre distintos instantes de la entrada. Las transiciones locales de corta duración son tratadas por las capas más bajas, mientras que las capas más altas relacionan información que involucra a períodos de tiempo más largos.

Redes de Jordan y Elman

Jordan [Moz92] estudió una clase de redes recurrentes, a veces denominadas redes secuenciales o redes de Jordan. Estas utilizaban un vector de estado que contenía copias de las activaciones de la capa de salida en el instante anterior, a su vez existían conexiones o pesos entre el vector de estado y la capa oculta. Como el vector de estado se puede ver como un vector de entrada extendido estas redes pueden entrenarse directamente por retropropagación. En la Figura 49 se observa el esquema de este tipo de redes.

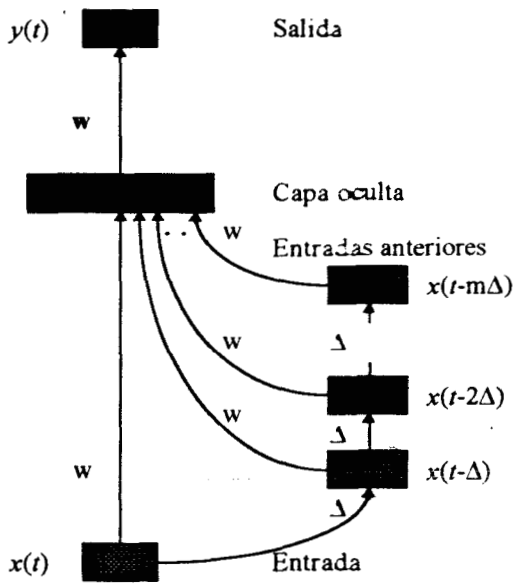

Figura 47: TDNN de una entrada y una salida

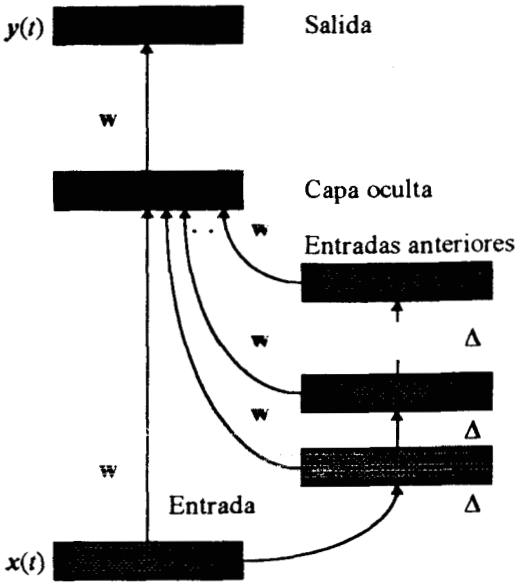

Figura 48: TDNN General

Elman [Elm89], [Elm90] introdujo una clase particular de redes recurrentes en la cual existe una conexión de realimentación entre un vector de estado y la capa oculta (Figura 50). Elman utilizó esta arquitectura, junto con el algoritmo de retropropagación, para aprender la estructura gramatical de un conjunto de oraciones generadas al azar a partir de un vocabulario limitado y una gramática. El vector de estado proporcionaba la aptitud de las redes de Elman y Jordan para almacenar información acerca de los estados anteriores (o salidas). Esto hizo pensar que si se utilizaban más vectores de estado se mejoraría el desempeño de las redes. Efectivamente esto lo que sucede en la mayoría de las aplicaciones [Wil93], [Wil95]. En la Figura 51 y la Figura 52 se observan los esquemas correspondientes, estas redes se denominan torres de Jordan y de Elman.

Retropropagación a través del tiempo

En las secciones anteriores se presentaron redes parcialmente recurrentes capaces de realizar asociaciones temporales. Este tipo de redes se denominan a veces redes recurrentes simples (SRNN). Sin embargo, en general, una red completamente recurrente es una alternativa más apropiada. En ese caso las unidades individuales pueden ser unidades de entrada, de salida, o ambas. Las salidas deseadas u objetivo son definidas sobre un conjunto arbitrario de unidades a determinados tiempos prefijados. Así mismo, pueden existir patrones de interconexión arbitraria entre las unidades. Un ejemplo de una red simple de dos unidades totalmente conectada se muestra en la Figura 53. La red recibe una secuencia de entrada $x(t)$ en la unidad 1, y se desea que la red genere la secuencia $d(t)$ como salida $y_2(t)$ de la unidad 2.

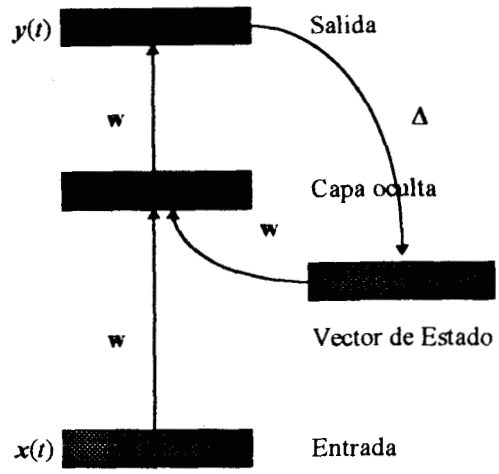

Figura 49: Red de Jordan

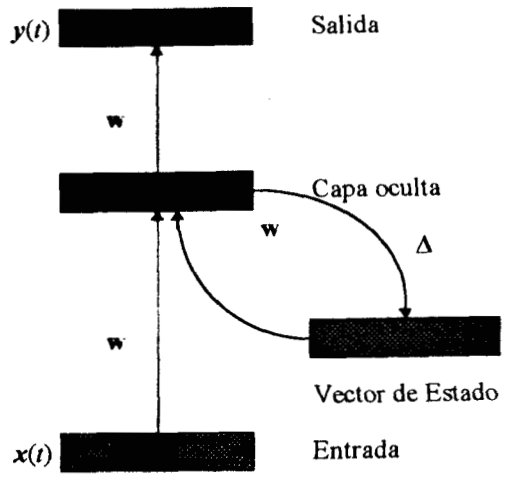

Figura 50: Red de Elman

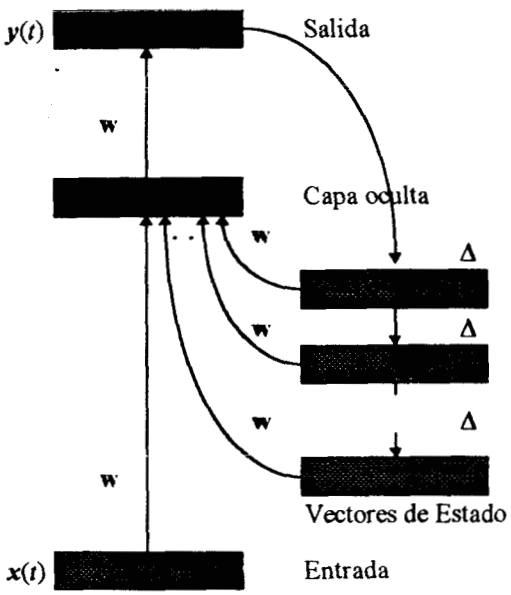

Figura 51: Torre de Jordan

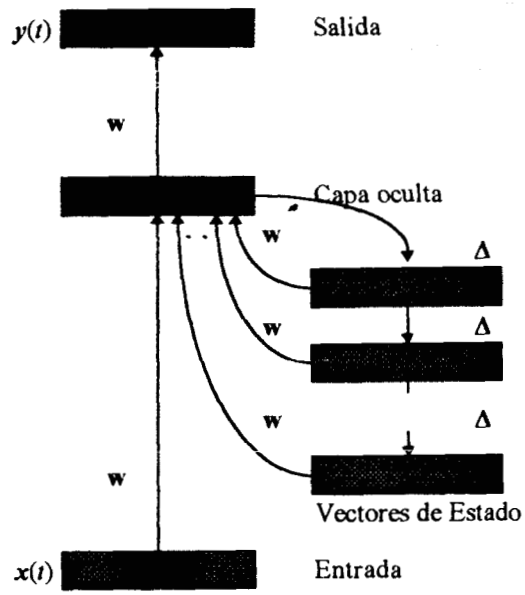

Figura 52: Torre de Elman

Una red que se comporta en forma idéntica a la red recurrente anterior sobre los tiempos *t=1, 2, 3* y *4* se muestra en la Figura 54. Esta proviene de 'desdoblar' la red recurrente en el tiempo [MiP69] para conseguir una red anteroalimentada multicapa. El número de capas resultante es igual al intervalo de tiempo de desdoblado *T*. Esta idea es efectiva si T es pequeño y limita la longitud máxima de las secuencias generadas. Aquí todas las unidades de la red recurrente son duplicadas *T* veces para que una unidad separada en la red desdoblada retenga el estado $y_i(t)$ de la red recurrente equivalente en el tiempo *t*. Se debe notar que las conexiones w_{ij} de la unidad j a la unidad y en la red desdoblada son idénticas para todas las capas.

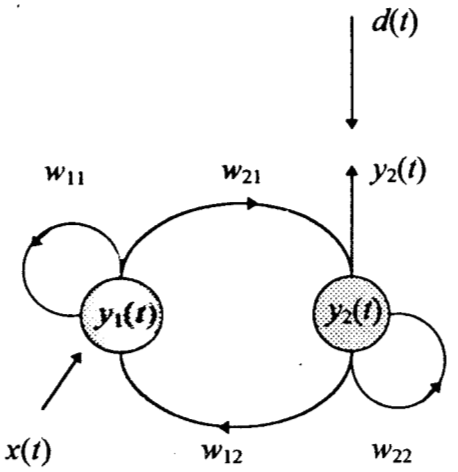

Figura 53: Red completamente recurrente

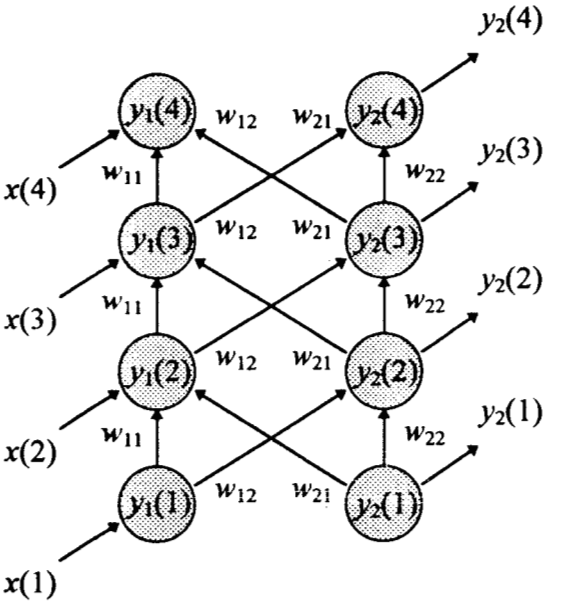

Figura 54: Red recurrente desdoblada

La red desdoblada resultante simplifica el proceso de entrenamiento para codificar la secuencia de asociación $x(t) \rightarrow d(t)$ porque se puede aplicar retropropagación. Sin embargo, se debe notar lo siguiente. Primero, se deben especificar objetivos para las unidades ocultas. De esta manera los errores a la salida de las unidades ocultas, y no sólo los errores de salida, deben ser propagados hacia atrás desde la capa en la cual se originaron. Segundo, es importante mantener la restricción de que todas las copias de cada peso w_{ij} se deben mantener idénticas a través de las capas duplicadas (retropropagación usualmente produce incrementos Δw_{ij} diferentes para cada copia particular del peso). Una solución simple es sumar los cambios individuales para todas las copias de un peso parcial w_{ij} y luego cambiar todas las copias con el Δw_{ij} total. Una vez entrenada los pesos de cualquier capa de la red desdoblada son copiados en la red recurrente, la cual va a ser usada para la tarea de asociación temporal. Adaptar retropropagación para trabajar con redes recurrentes desdobladas resulta en un algoritmo denominado retropropagación a través del tiempo [RHW86]. Existen relativamente pocas aplicaciones de esta técnica en la literatura [RHW86], [Now88], [NgW89]. Una razón es su ineficiencia en manejar secuencias largas. Otra razón es que existen otros métodos de aprendizaje que pueden resolver el problema sin desdoblar la red. Entre estos métodos se encuentran Retropropagación Recurrente, Retropropagación Recurrente Dependiente del Tiempo y Aprendizaje Recurrente en Tiempo Real. La complejidad de estos métodos es mucho mayor que las simplificaciones mencionadas y poseen algunos problemas con los mínimos locales. Para una revisión de estos métodos ver [Hau95] y [TNN94].

Criterios para la elección de la Arquitectura Neuronal

En este trabajo se requiere utilizar una arquitectura que aproveche las ventajas de los análisis realizados por la STFT y DWT. Se pueden enumerar una serie de características deseables a los fines de orientar la búsqueda [GWS92], [WHH89], [Tak95]:

- Invariabilidad a la translación temporal: Esto quiere decir que la Red debe ser capaz de aprender y reconocer patrones a pesar de corrimientos temporales de los mismos.

- La red debe poseer una cantidad suficiente de capas y de interconexiones entre las unidades correspondientes a cada una de esas capas. Esto permitirá que la misma discrimine regiones de decisión no lineales complejas.

- El número de pesos ajustables de la red deberá ser suficientemente pequeño comparado con la cantidad de datos de entrenamiento, de tal modo que la red este obligada a extraer regularidades modificando los pesos de las interconexiones.

- La red debe poder representar internamente las relaciones entre eventos no simultáneos, aunque próximos en el tiempo. Es decir que la red debe poder aprender la dinámica de la señal.

- La estrategia de entrenamiento no debe requerir una alineación temporal precisa de los patrones a ser aprendidos. Para ello, las representaciones internas realizadas por la red, deben ser invariantes a la traslación en el temporal.

- El tiempo requerido para el entrenamiento y prueba de la Red debe ser razonable en relación a la cantidad de datos y el hardware disponible.

- En lo posible se utilizará software comercial disponible o previamente desarrollado para el entrenamiento de la Red.

Recientemente se han empleado diversas arquitecturas de redes (en especial TDNN's) en conjunción con Wavelets, e inclusive con HMM's aplicados al reconocimiento de voz con buenos resultados [Fav94].

De acuerdo con estos criterios y con los resultados de algunas pruebas preliminares en nuestro trabajo se utilizaron redes recurrentes simples, en especial las TDNN y las de Elman. Por otra parte no es el objetivo de esta tesis encontrar el mejor clasificador y como la cantidad de experimentos y datos es relativamente grande esta elección estuvo principalmente orientada por los dos últimos criterios.

Redes de Kohonen

Los Mapas Auto Organizativos o redes de Kohonen son un tipo de redes que emplean ANS y que por lo tanto poseen la capacidad de aprender sin necesidad de que los ejemplos estén

etiquetados. Estas redes son capaces de separar los datos en un número especificado de categorías o clases.

En este caso solo existen dos capas (Figura 55) : una capa de entrada y una capa de salida. La capa de entrada tiene tantas neuronas como atributos. En la capa de salida existe una neurona para cada posible categoría de salida, el número máximo de categorías se especifica de antemano. Los valores de activación de salida se calculan de acuerdo a la distancia entre el patrón de entrada y los pesos de cada neurona.

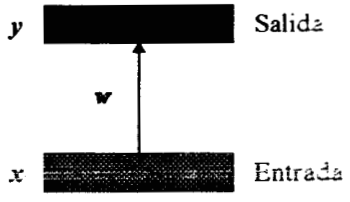

Figura 55: Red de Kohonen

Los patrones de entrenamiento son presentados a la capa de entrada, luego propagados a la capa de salida y evaluados. Una sola neurona de salida es la "ganadora" en función de su nivel de activación. Los pesos de la red son ajustados durante el entrenamiento. Este proceso se repite para todos los patrones. Esta red es muy sensible a la tasa de aprendizaje, la cual decae exponencialmente a medida que progresa el entrenamiento produciendo cada vez menores cambios en los pesos. Esto causa que la red llegue a un punto de estabilización donde culmina el entrenamiento.

La red ajusta los pesos para todas las neuronas en una vecindad alrededor de la neurona ganadora. El tamaño de esta vecindad también es variable, comenzando en valores grandes (inclusive alcanzando el número total de categorías) y decreciendo hasta cero al final del entrenamiento, con lo que solo los pesos de la neurona ganadora son actualizados. En ese momento también la tasa de aprendizaje es pequeña y los clusters o agrupamientos ya están definidos. De esta manera al finalizar el entrenamiento los pesos de cada neurona corresponden al centroide de la clase o agrupamiento correspondiente por lo que se puede utilizar esta red para encontrar estos centroides.

Elección de la Familia de Wavelets

Según se mostró en el capítulo anterior no existe ningún criterio analítico para la elección de una familia de wavelets. Además esta elección depende fuertemente de la aplicación. Por ejemplo una familia de Wavelets que funciona muy bien en la compresión de algún tipo de señales, puede que no funcione tan bien en un problema de clasificación de las mismas. Una alternativa es la experimentación exhaustiva con cada posible familia y cada posible parámetro de cada familia. Existen dos inconvenientes para nuestra aplicación de este último procedimiento : uno es la gran cantidad de tiempo requerido para cada entrenamiento, el otro es que dependiendo de una serie de factores los resultados dados por la red pueden variar para el mismo experimento en forma considerable. La idea entonces sería emplear alguna técnica o criterio que nos permita realizar la elección pero sin tener que probar o entrenar una red para cada familia y para cada juego de parámetros. En esta sección se presenta un método alternativo.

En la Figura 56 se puede apreciar un esquema del proceso sugerido para la selección de una familia o base de wavelets. En primera instancia se genera un archivo de entrenamiento con

los frames de la señal de voz sin procesar y sus correspondientes etiquetas, este se separa en cinco archivos, uno para cada clase de fonemas. De cada archivo se obtienen 3 centroides para cada clase. Todo esto esta considerado en nuestro primer bloque del esquema (que podría ponerse también como cinco bloques en paralelo). A continuación se calcula la DWT de cada centroide con la familia en consideración. Una vez obtenidos los centroides en el dominio transformado se procede a calcular las distancias entre estos como una medida de la separación de las clases correspondientes en este dominio. Finalmente se calcula la distancia total entre centroides y la distancia mínima, que son nuestros indicadores finales. Este proceso se repite con cada familia en consideración. Con estos valores para cada familia se procede a escoger la familia que posee mayor distancia total, en el caso que existan dos familias con valores similares entonces se elige aquella con mayor distancia mínima entre clases.

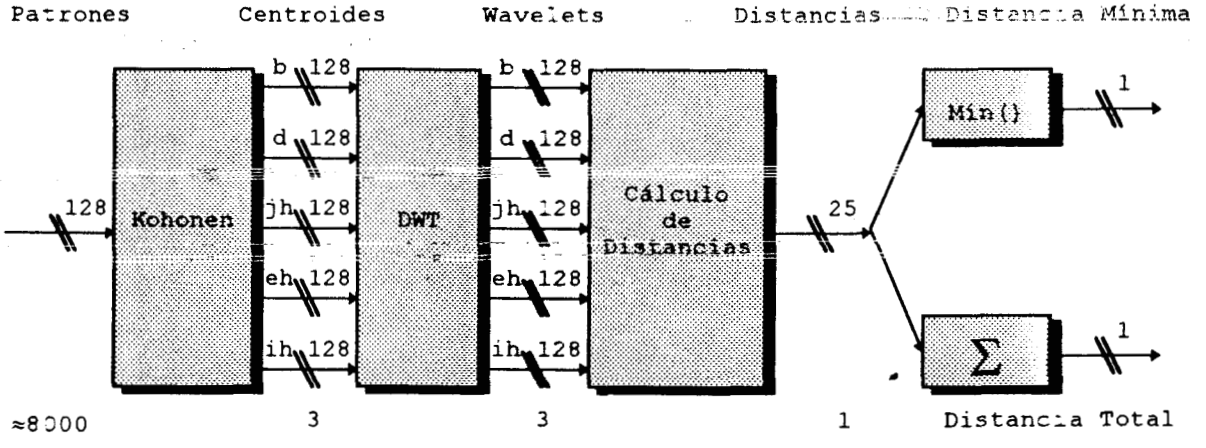

Figura 56: Esquema del Proceso de Selección de la Base

En la Figura 57 se observa un diagrama simplificado de distribución de clases en un espacio bidimensional hipotético. Aquí se puede apreciar mejor como cuando las distancias entre los centroides de las clases son mayores se facilita la clasificación, además la distancia mínima constituye un índice para la clase con más probabilidad de confusión.

Existen varias medidas de distancia que se pueden emplear. En nuestro caso primero normalizamos la energía de los coeficientes de la transformación wavelet (wc) :

$$p = \left(\frac{wc}{|wc|} \right)^2$$

Esta normalización es importante si existe gran diferencia de energía entre los patrones. Luego de esta normalización los coeficientes p para los 3 centroides de la clase i se acumulan en p_i. Finalmente se calcula la distancia euclídea entre cada para de clases i, j de la siguiente forma:

$$D(i,j) = \left| p_i - p_j \right|^2$$

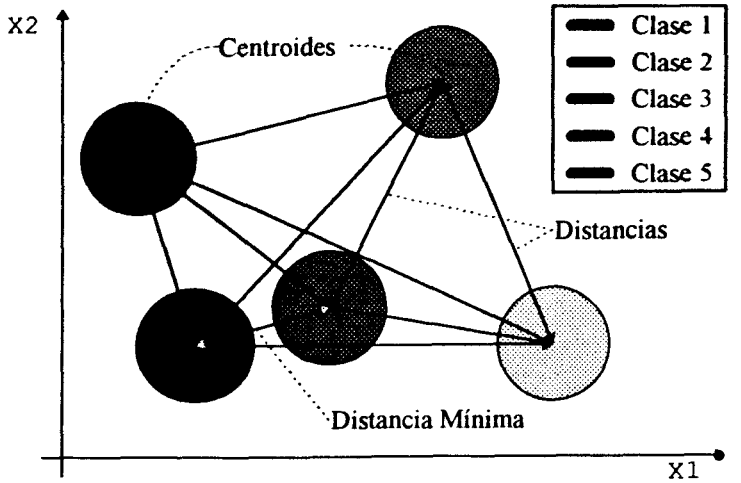

Figura 57: Diagrama Simplificado de Distribución de Clases

La decisión de tomar 3 centroides para cada clase se tomó como un compromiso entre complejidad o tiempo empleado por el método y efectividad del mismo. Un solo centroide constituía a nuestro parecer una sobre-simplificación que podría afectar los resultados. A partir de un análisis de los centroides encontrados, se vio que se distribuyeron principalmente en diversos rangos de energía de los patrones. Esto quiere decir que en todas las clases se encontró un centroide correspondiente a un grupo de alta energía, otro de energía media, y otro de pequeña energía.

Aspectos de Implementación Práctica

Durante las comparaciones se intentó mantener el número total de pesos y umbrales constante, de manera de que la estructura no fuera un factor que pesara en la diferencia de desempeño entre las distintas alternativas de preproceso (distintas Wavelets y Fourier). Todas las técnicas implementadas generan patrones de 128 dimensiones o coeficientes. Esto produce una tasa de 1 frame o patrón cada 8 mseg. Como las redes utilizadas son recurrentes en general la información que procesan por unidad de tiempo corresponde a varios frames (el actual y uno o dos anteriores). La cantidad de salidas corresponde a la cantidad de clases o fonemas a clasificar, que en este caso son 5 (/b/, /d/, /jh/, /ih/, /eh/).

En todos los casos se ajustaron pesos y umbrales con el algoritmo de retropropagación con momento. Los parámetros se mantuvieron fijos durante el aprendizaje. Luego de bastante experimentación los siguientes valores parecieron adecuados. La tasa de aprendizaje se estableció en 0.1, el factor de momento también en 0.1 y los pesos iniciales aleatorios entre -0.3 y 0.3. Entradas y salidas se normalizaron entre 0 y 1 utilizando los valores máximos y mínimos correspondientes pero extendiendo el rango en un 5 % con respecto al archivo de entrenamiento.

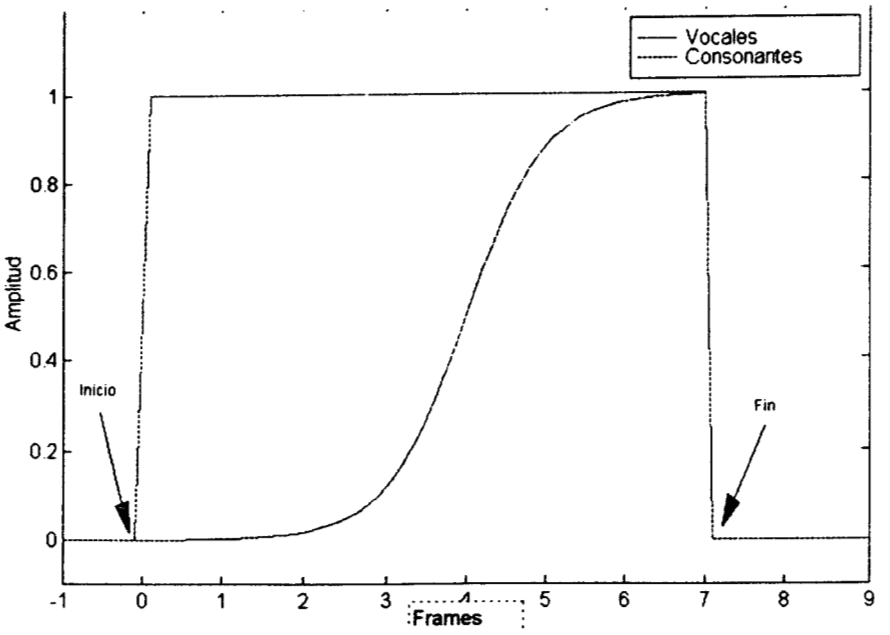

Figura 58: Curvas de Activación Deseadas para Vocales y Consonantes

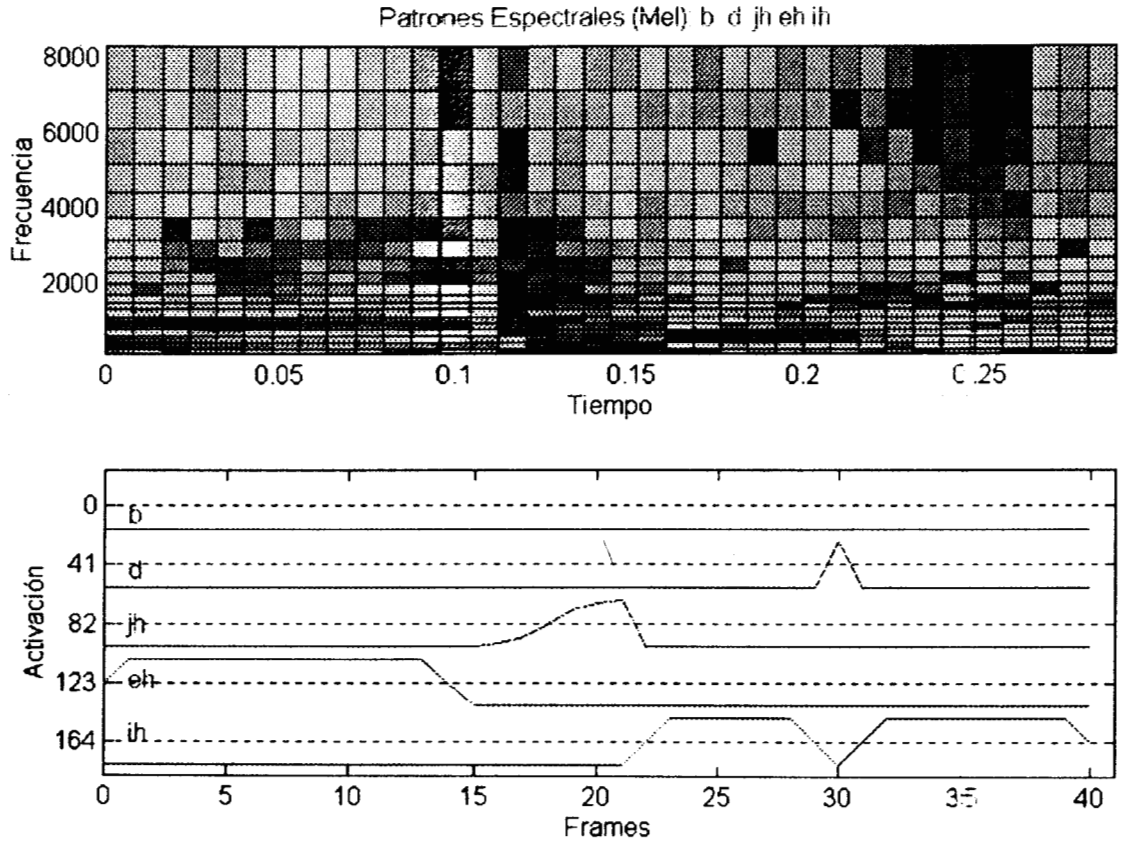

Figura 59: Patrones espectrales y Curvas de Activación deseada (s= 1)

Como se mencionó en el capítulo que trata sobre los datos el tamaño de los archivos y el tiempo necesario para el entrenamiento hicieron prácticamente imposible la utilización de un método más preciso de estimación de la capacidad de generalización de las redes. Por esta razón se empleó el método más sencillo de separar los datos en un conjunto de entrenamiento y otro de prueba. Para ello se utilizó la partición de los datos que ya viene en la base TIMIT y que obedece a criterios que pretenden aseguran que los datos de prueba sean representativos de las clases implicadas mientras que sean distintos a los de entrenamiento.

La forma de las salidas deseadas utilizadas para vocales y consonantes se puede apreciar en la Figura 58. La razón de esta diferencia es que en el caso de las vocales el espectro se mantiene relativamente constante durante la emisión, por lo que bastan uno o dos frames (espectros o wavelets) para identificar el fonema. Sin embargo, en el caso de las consonantes solo se puede emitir una clasificación segura después de que se ha seguido la evolución espectral de la emisión durante más tiempo. En la Figura 59 se observa un ejemplo.

Se sabe que si se la red se expone demasiado a los datos de entrenamiento se produce un fenómeno de sobreaprendizaje que afecta la capacidad de generalización de la misma. Este fenómeno se puede observar en las curvas de error que aparecen en la Figura 60 para un caso hipotético. Para evitar este efecto en todos los experimentos el aprendizaje se detuvo en el pico de generalización. Esto puede hacer parecer que los resultados con respecto a entrenamiento no son tan buenos como los esperados. El error con respecto al archivo de prueba se calculó cada 2000 iteraciones del algoritmo (aproximadamente 4 veces por época del archivo de entrenamiento).

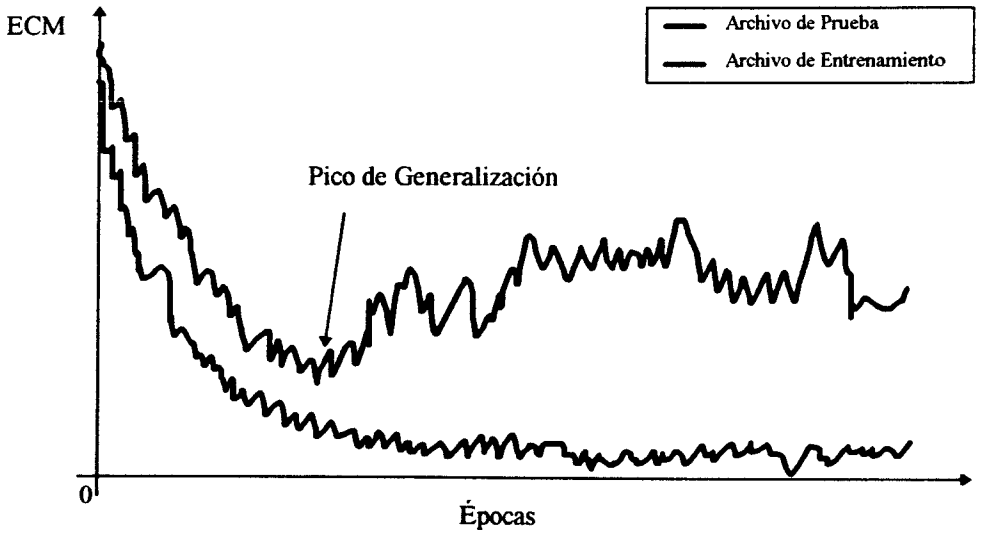

Figura 60: Sobreaprendizaje

En el caso de la red de Kohonen empleada para calcular los centroides el número de entradas fue otra vez 128 (muestras de la señal de voz sin procesar). Las salidas para cada una de las 5 redes entrenadas fueron 3, para obtener 3 centroides para cada clase. La tasa de aprendizaje inicial fue en todos los casos de 0.99, la vecindad inicial de 2, el valor inicial de los pesos aleatorio entre -0.1 y 0.1, y el algoritmo se corrió por 500 épocas.

VI . Resultados y Conclusiones

Introducción

El objetivo principal de este trabajo consiste en comparar la STFT con la DWT como etapas de preproceso de un sistema de RAH. Como en el caso de DWT existen muchas posibilidades que afectan este preproceso (distintas familias, parámetros, agrupamiento de los coeficientes, etc.), esta comparación no estaría completa si no se exploraban algunas de estas alternativas para encontrar la mejor en nuestro caso.

En los capítulos anteriores hemos presentado el problema de RAH, explicado los aspectos fisiológicos relevantes, y analizado las principales características de la señal de voz. Además, hemos presentado los datos con los que se realizaron los experimetos (TIMIT) y definido las propiedades de Fourier y Wavelets, necesarias para comprender y utilizar estas herramientas para generar los patrones de entrenamiento. Para la etapa de clasificación se mostraron las distintas posibilidades y se expusieron los criterios para la elección del tipo de clasificador empleado (Redes Neuronales Recurrentes). Nos resta aquí una breve explicación acerca de los experimentos realizados y la presentación e interpretación de los resultados obtenidos.

Experimentos Realizados

En primera instancia se procedió a aplicar el método explicado en el capítulo anterior para la elección de los parámetros de las diferentes familias. Este método se aplicó para cada familia en forma separada, variando los parámetros correspondientes (cantidad de momentos, regularidad, etc.). Un esquema del proceso se puede apreciar en la Figura 61.

Una vez conseguidos los valores óptimos de los parámetros se procedió a repetir el proceso pero ahora para elegir la mejor familia de wavelets. El proceso se esquematiza en la Figura 62. Estos resultados se compararon posteriormente con los obtenidos por medio de las redes neuronales para verificar la validez del criterio.

Finalmente se entrenaron las redes para Fourier y Wavelets (solo con los parámetros óptimos) como se muestra en la Figura 63. También se hicieron una serie de pruebas preliminares con los patrones obtenidos con Fourier en escala de Mel. Los resultados se presentan en términos de los porcentajes de reconocimiento con respecto a los archivos de entrenamiento y prueba. Las matrices de confusión se obtienen a partir de las diferencias para cada clase entre la clasificación dada por la red y la real. Estas matrices permiten analizar mejor los conflictos entre los fonemas más dificiles de clasificar. Esto nos sirve para guiar los experimentos y establecer conclusiones más "finas" que con los índices globales.

Para los experimentos basados en redes neuronales se empleó el programa comercial NeuroShell [Neu2.3] que provee una gran variedad de arquitecturas y entre ellas varias de tipo recurrente. Para la generación de los patrones de entrenamiento, a partir de los datos de

TIMIT, se confeccionaron una serie de programas en Matlab, con la ayuda de el toolbox de Procesamiento de Señales, el toolbox de Wavelets de la Universidad de Vigo (Uvi_Wave) [GoG95] y el toolbox de Wavelets de la Universidad de Stanford (WaveLab) [BCD95], [BCD95b]. Estos programas leían los fonemas elegidos para las regiones y hablantes seleccionados y creaban un archivo con los coeficientes del análisis, los valores de activación deseados y la etiqueta del fonema correspondiente. Además producian una serie de gráficos y archivos auxiliares de estadísticas para control del proceso. Para la generación de algunas gráficas también se utilizó el toolbox SPC de Análisis de Señales [Brow95] y software desarrollado previamente para análisis de voz (en lenguaje Pascal) [ARZ93], [Aru94], [ARZ94].

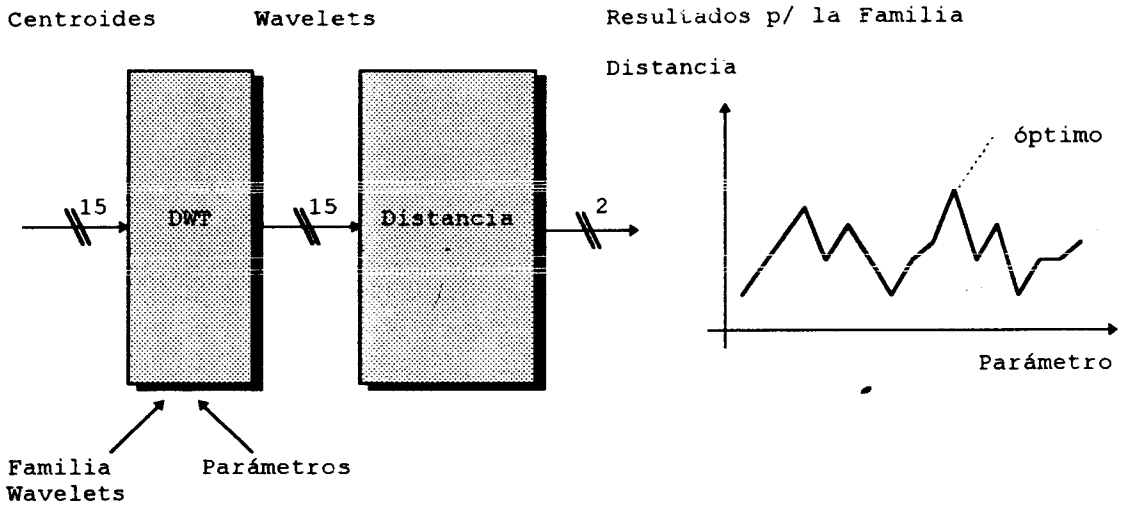

Figura 61: Selección de los Parámetros óptimos para c/ Familia

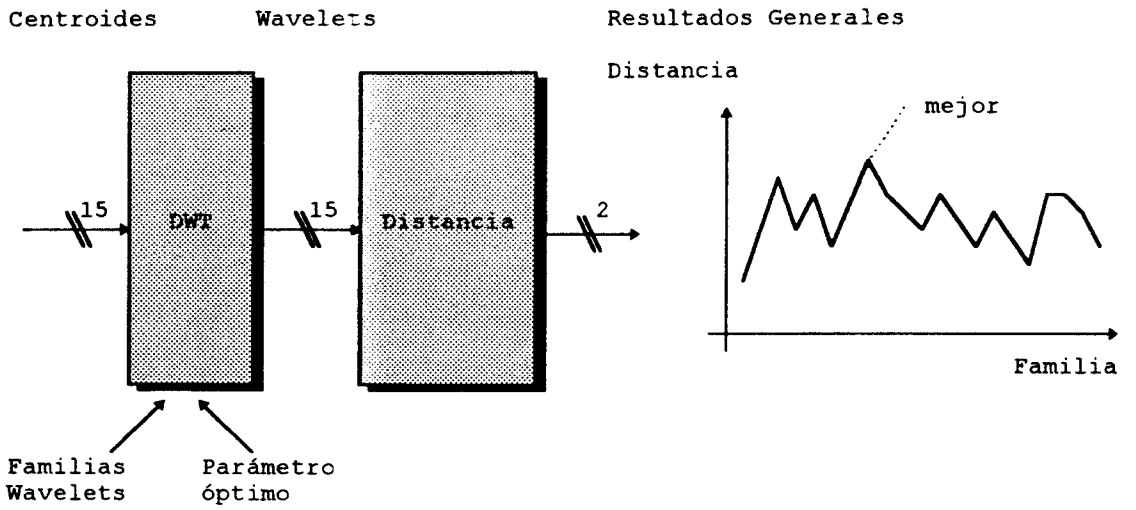

Figura 62: Selección de la Mejor Familia

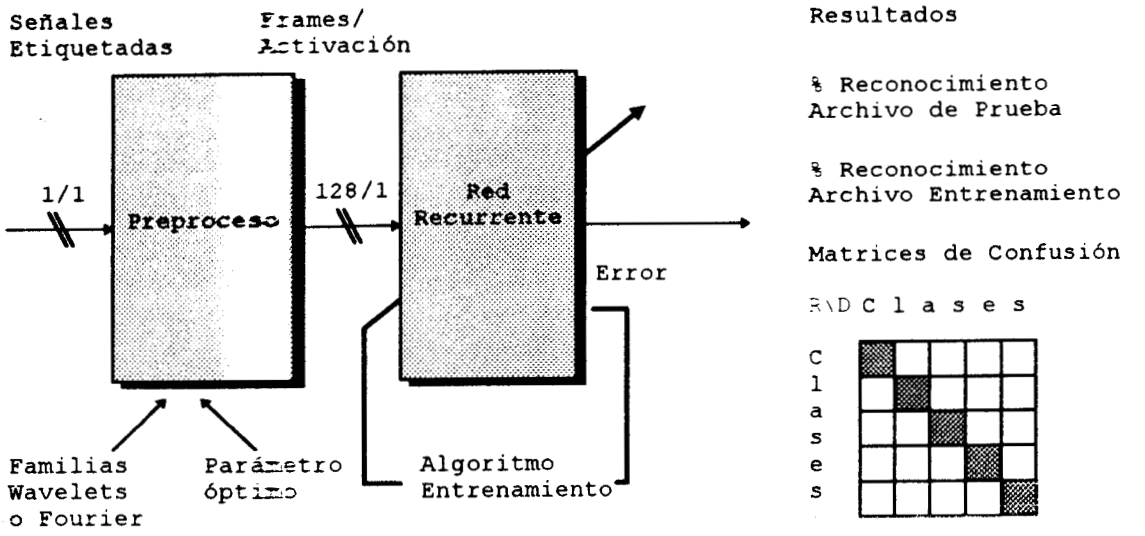

Figura 63: Entrenamiento de las Redes Recurrentes

Resultados

Los experimentos preliminares con redes se corrieron con los patrones generados por la STFT en escala de Mel. Esto se debe a que, por tener solo 20 coeficientes por cada frame, la estructura de las redes ensayadas resultaba más pequeñas y los tiempos de entrenamiento menores. Como ya se indicó al principio se incluyó en los archivos de entrenamiento la región oculsiva de /b/, /d/ y /jh/. Sin embargo se observó en las matrices de confusión correspondientes que la mayoría de los conflictos entre estas clases se debían a que a nivel acústico no existían diferencias en esta porción (inclusive en TIMIT poseen el mismo símbolo). Por eso el resto de los experimentos se llevó a cabo sin esta porción de los fonemas (que puede tomarse como otra clase).

En la Tabla 8 se pueden ver los resultados para las distintas arquitecturas ensayadas. En este caso la red de Elman empleada posee una pequeña modificación que le permite acumular más de un instante en el vector de estado. Esto se logra pesando el valor anterior con un factor de memoria y sumándolo al actual. En el caso de la TDNN los instantes anteriores se presentan como las derivadas o diferencias. Como se puede apreciar la arquitectura que mejor funcionó fue la TDNN por lo que fue la seleccionada para el resto de los experimentos. En la Figura 64 se muestran las salidas deseadas y las reales para la TDNN ya entrenada.

Descripción	Estructura	Entrenamiento (%)	Prueba (%)	Épocas
TDNN	20+20+20/135/5	82.56	81.83	98
Elman	20+80/80/5	78.02	77.07	122
Elman+TDNN	20+20+20+87/87/5	68.71	67.12	105

Tabla 8: Resultados para Fourier en Escala de Mel

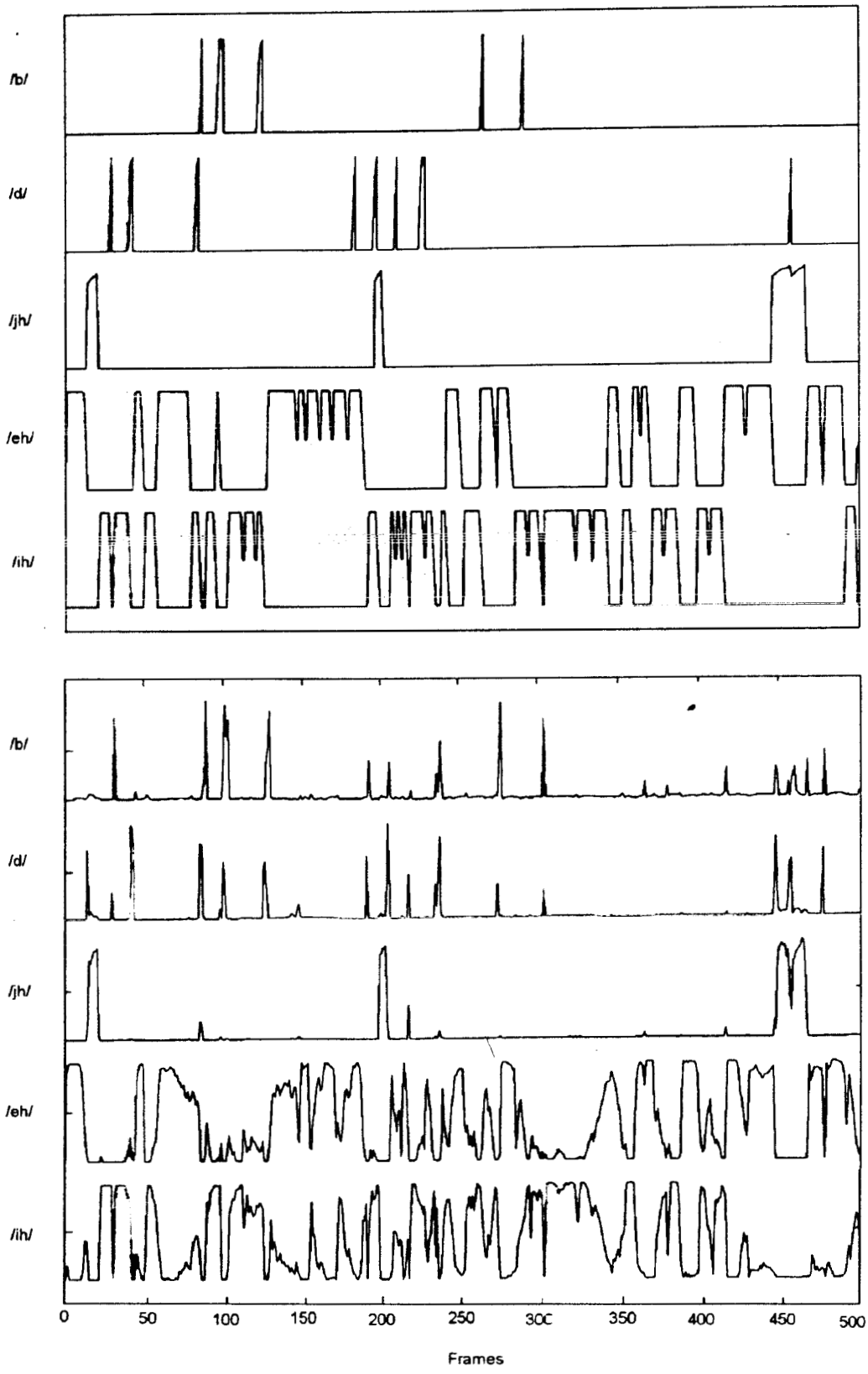

Figura 64: Salidas Deseadas y Salidas de la Red para cada fonema

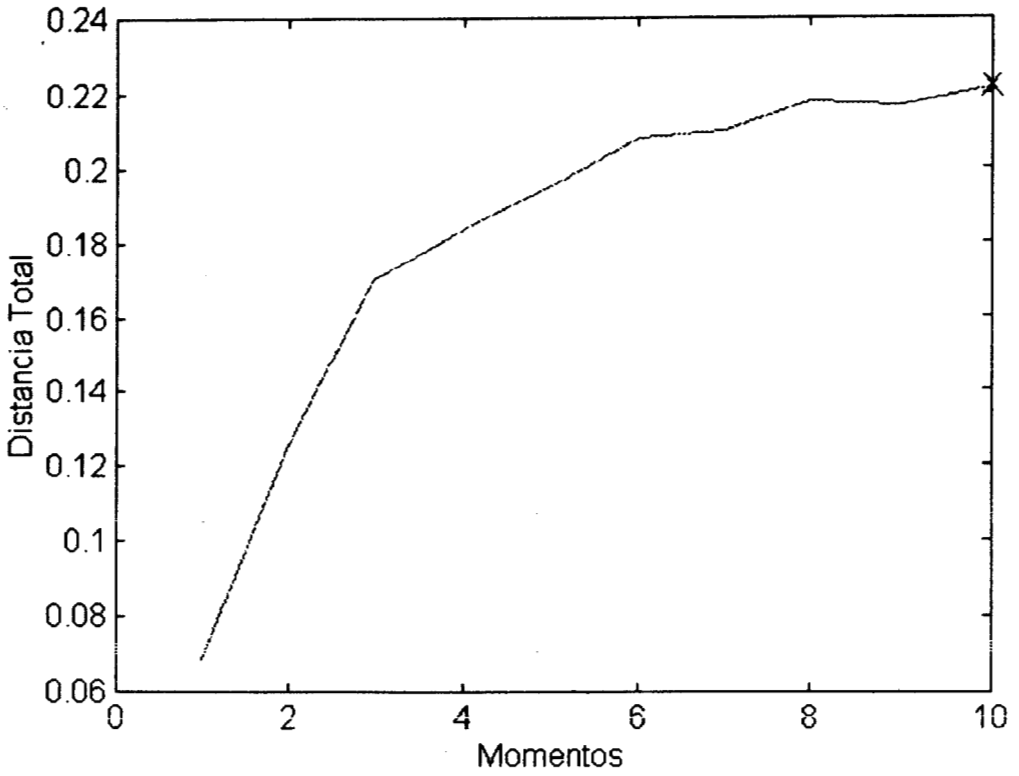

Figura 65: Comparación de Wavelets Daubechies (Dist. Total)

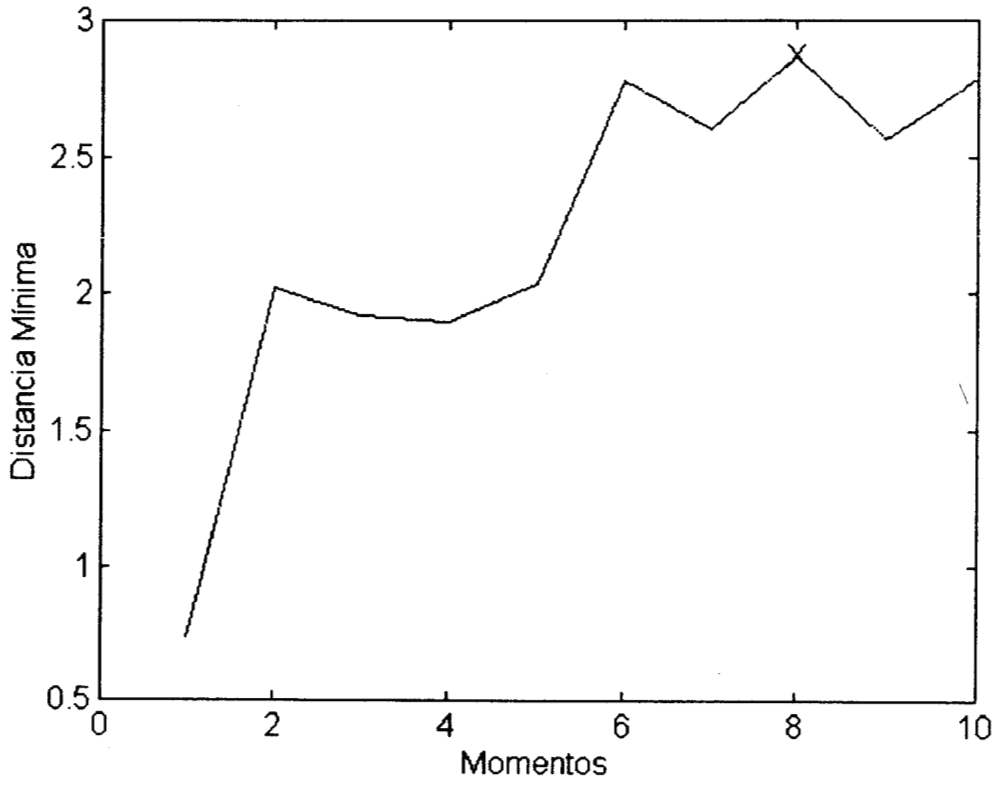

Figura 66: Comparación de Wavelets Daubechies (Dist. Mínima)

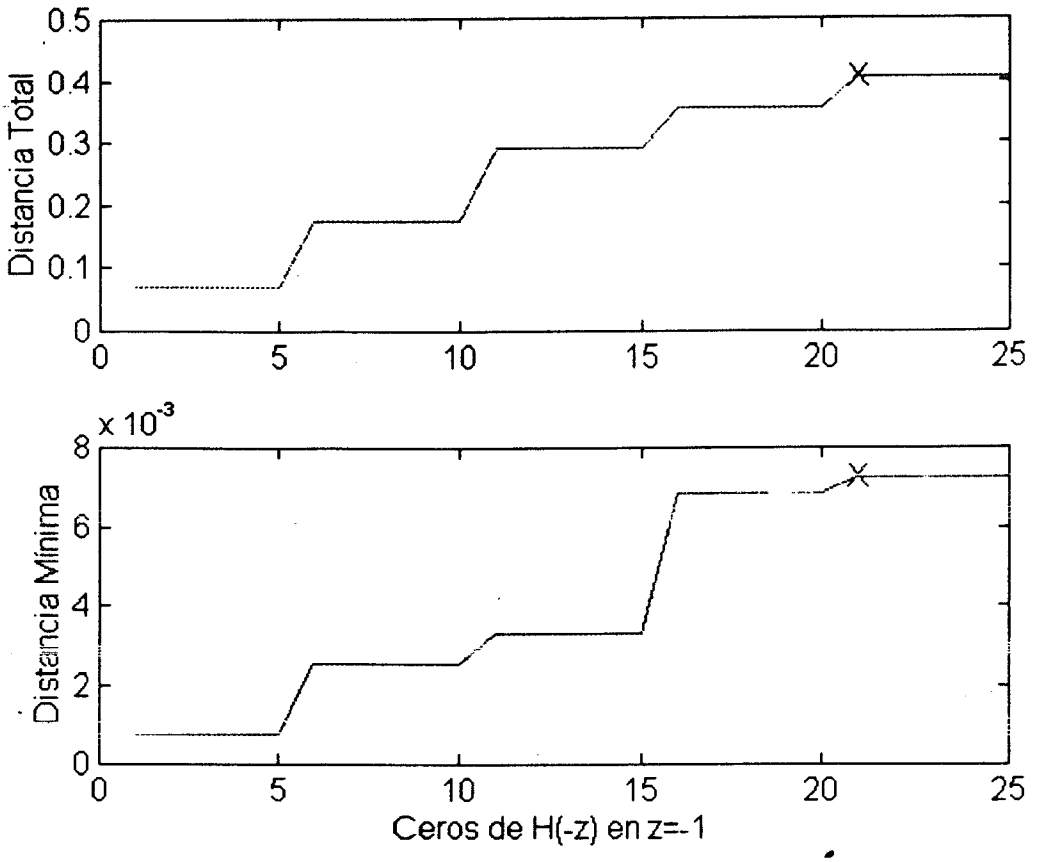

Figura 67: Comparación de Wavelets Splines (m)

Luego presentamos los resultados arrojados por el criterio para selección de la familia de wavelets. Este se utilizó para elegir los parámetros adecuados para cada familia (en los casos de Daubechies y Splines, donde existen parámetros). En la Figura 65 y la Figura 66 se muestran las gráficas correspondientes al caso de Daubechies. Aquí el valor seleccionado fue de 8 momentos (m=16). En la Figura 67 se observan los resultados para el caso de Splines (se juntaron ambos parámetros en una sola dimensión). El resultado corresponde a m=9 como óptimo, el parámetro n no influyó en los resultados por lo que se escogió como n=37 de manera que la wavelet obtenida fuera regular.

Familia	Distancia Total	Distancia Mínima
Splines (9,37)	0.5152	0.0150
Meyer	0.3946	0.0058
Daubechies (16)	0.2179	0.0028
Vaidyanathan	0.2174	0.0032
Haar	0.0685	0.0007

Tabla 9: Discriminación de los centroides con diferentes wavelets

Posteriormente se procedió de igual forma para elegir la familia según se explicó en la sección anterior. En la Tabla 9 se presentan los resultados correspondientes.

A continuación se presentan los resultados correspondientes a las pruebas definitivas con las TDNN de estructura fija. La estructura se derivó de los mejores resultados en las pruebas hechas con la STFT (128 coeficientes) y se mantuvo fija en el resto de los experimentos. Todas se pararon en el pico de generalización y se corrió por lo menos dos veces cada experimento partiendo de diferentes pesos eligiéndose el mejor valor para la tabla.

Descripción	Estructura	Entrenamiento (%)	Prueba (%)	Épocas
Fourier	128+128/150/5	79.67	77.53	45
Splines (9,37)	128+128/150/5	70.43	70.92	75
Vaidyanathan (*)	128+128/150/5	69.25	68.11	120
Meyer	128+128/150/5	65.46	67.58	50
Daubechies (16)	128+128/150/5	63.76	63.11	163
Haar	128+128/150/5	53.15	50.00	178

Tabla 10: Porcentaje Reconocimiento de las Redes

Se debe aclarar que en el caso de Vaidyanathan los resultados iniciales fueron cercanos a los de Haar, con la salida fija en alto en una de las vocales. Este problema tenía su origen en la diferente distribución de los fonemas (ver capítulo sobre los datos) y para solucionarlo fue necesario la implementación de un entrenamiento en dos etapas. En la primera se presentó solo una parte del archivo original, de tal manera que existiera aproximadamente igual cantidad de ejemplos de cada clase. Cuando se alcanzó el pico de generalización se procedió a entrenar con el archivo completo, otra vez hasta el pico de generalización.

Para proseguir presentamos las matrices de confusión con respecto a los datos de prueba para todos los experimentos realizados.

	R/D	b	d	jh	eh	ih
Fourier	b	52.6	6.8	0.0	0.5	0.2
	d	18.4	63.6	2.8	0.0	0.0
	jh	0.0	2.3	97.2	0.0	0.0
	eh	21.0	6.8	0.0	83.6	28.1
	ih	7.9	20.4	0.0	15.8	71.7

	R/D	b	d	jh	eh	ih
Meyer	b	65.8	8.0	0.0	0.2	0.3
	d	19.5	61.3	12.5	0.2	0.7
	jh	0.0	4.0	85.2	0.1	0.0
	eh	2.4	5.3	0.0	43.3	14.3
	ih	12.2	21.3	2.3	56.1	84.7

	R/D	b	d	jh	eh	ih
Daub.	b	44.9	5.7	0.0	0.1	0.0
	D	36.7	63.2	30.7	0.9	2.3
	jh	0.0	1.1	67.1	0.1	0.0
	eh	4.1	6.9	0.0	50.6	22.0
	ih	14.3	22.9	2.3	48.3	75.6

	R/D	b	d	jh	eh	ih
Haar	b	30.8	18.0	0.0	0.0	0.3
	d	0.0	8.0	0.0	0.0	0.0
	jh	0.0	32.0	96.5	0.2	0.2
	eh	42.3	38.0	3.5	97.5	96.2
	ih	26.9	4.0	0.0	2.2	3.2

	R/D	b	d	jh	eh	ih
Vaid.	b	33.3	8.2	1.2	0.4	0.7
	d	41.7	63.0	15.7	0.6	1.5
	Jh	0.0	12.3	81.9	0.0	0.1
	eh	8.3	4.1	0.0	68.9	29.7
	ih	16.7	12.3	1.2	30.1	68.0

	R/D	b	d	jh	eh	ih
Splines	b	47.0	3.8	0.0	0.0	0.2
	d	34.4	78.9	17.2	0.0	0.2
	jh	0.0	3.8	82.7	0.1	0.1
	eh	12.5	7.6	0.0	67.2	25.9
	ih	6.2	5.7	0.0	32.6	73.5

Tabla 11: Matrices de Confusión (Archivo de Prueba)

Interpretación y Conclusiones

Los resultados reportados en este trabajo constituyen del orden de 1000 horas máquina corriendo en una computadora Pentium a 100 MHz y código optimizado. Esta enorme carga de computo a imposibilitado realizar todas las pruebas previstas, limitándonos a las más importantes.

Pese a nuestras expectativas iniciales Fourier se comportó generalmente de manera más robusta y con mejor capacidad de generalización que la implementación y las familias de wavelets ensayadas. Esto quizas se deba a una mejor resolución frecuencial de Fourier a frecuencias medias y altas, contrastando con una baja resolución de Wavelets a estas frecuencias y una excesiva resolución temporal. Esto se puede apreciar en las matrices de confusión de las wavelets por los altos valores de confusión entre las vocales debido a que la resolución en frecuencia no alcanza para distinguir las pequeñas diferencias entre las formantes (ver el criterio de elección de las dos vocales en el capítulo sobre los datos). El tamaño de la ventana de análisis también puede influir en los resultados y es un parámetro que aquí no se modificó. En un trabajo reciente [RHPF] se compararon una serie de preprocesos (FFT, bancos de filtros, modelos de oído y LPC) para un sistema de RAH sobre TIMIT con redes recurrentes. En este trabajo no se encontró una diferencia significativa entre los distintos análisis, aunque no se experimento con wavelets; y se concluyó que eran mucho más significativos los cambios de arquitectura o estructura de las redes. También

hemos citado aquí uno de los pocos trabajos de RAH con wavelets encontrados [Fav94]. En este trabajo se realiza también una comparación con Fourier y los resultados son favorables para Wavelets. Sin embargo existen varias diferencias entre este trabajo y el nuestro. Primero la tarea de reconocimiento es el E-set de TI-46, que se puede considerar más sencilla que TIMIT. Además en [Fav94] se trabaja con la Transformada Wavelet Continua Muestreada (TWCM) y se muestrean los coeficientes en escala de Mel, mejorando la resolución en frecuencia con respecto a nuestro caso (6 coeficientes por octava contra 1). Este enfoque requiere más tiempo de cálculo que el algoritmo rápido. También ensaya diferentes formas de agrupar los coeficientes (en nuestro solo se pudo emplear uno por el problema de los tiempos de entrenamiento).

De las matrices de confusión se puede ver que el fonema que se identificó mejor en promedio (o que se confundió menos) fue la /jh/. Esto se debe a la banda de energía de alta frecuencia que permite al clasificador separarlo rápidamente del resto.

De las familias de Wavelets utilizadas la que mejor se comportó fue Splines. Podemos también ver que los porcentajes de reconocimiento estan distribuidos en forma bastante pareja entre las clases. La peor, como era de esperarse, fue Haar debido probablemente a su mala localización tiempo-frecuencia. Aunque pareciera que las wavelets simétricas (Splines y Meyer, véase Tabla 12) tuvieron resultados más uniformes no se puede asegurar una relación. No se advierte una relación con la existencia o no de soporte compacto o la regularidad (el método de selección no encontró diferencia en este caso).

Familia	Soporte Compacto	Simetría	Regularidad	Localización	Comentario
Haar	SI	SI	NO	mala	la más simple
Meyer	NO	SI	SI	buena	muy difundida
Daubechies	SI	NO	variable	variable	optimiza suavidad
Splines	SI	SI	variable	variable	biortogonal
Vaidyanathan	SI	NO	SI	buena	codificación de voz

Tabla 12: Resumen de características de las Wavelets

Obsérvese la gran correlación entre las posiciones relativas de las familias dadas por el criterio propuesto para la selección y los resultados reportados para las redes . Hay que tener en cuenta que los resultados para Vaidyanathan se obtuvieron de manera distinta al resto, lo que sugiere que a veces las redes pueden hacer un mal trabajo en elegir un método frente a otro. Esto se debe a que puede existir un conjunto de pesos que resuelva muy bien un problema pero que el algoritmo de entrenamiento no lo encuentre (por ejemplo si existen gran cantidad de mínimos locales con valor de error alto). Mediante el método de selección de la base wavelet propuesto también se puede generar una matriz de distancias para cada familia. Esta mostró tener gran correlación con las matrices de confusión correspondientes y podría usarse como sustituto de estas. Aunque hace falta experimentar más con el criterio de selección, todo esto sugiere que puede ser muy útil en aplicaciones de clasificación. En

particular como alternativa a la costosa (y a veces poco confiable) experimentación exhaustiva con las redes recurrentes. La falta de este tipo de criterios en la literatura aumenta la importancia práctica de este resultado.

Recomendaciones y sugerencias finales

El problema de la resolución insuficiente en algunos casos podría resolverse con un esquema basado en Wavelets Packets. Características como soporte compacto, regularidad, etc., no parecen decir nada acerca de como se comporta una familia en determinada aplicación. El método propuesto aquí no es la única alternativa. Otras técnicas emparentadas con Wavelets, como Matching Pursuit, podrían constituir otra herramienta de análisis.

No se han utilizado las wavelets basadas en modelos de oído que podrían representar un enfoque todavía más fisiológico y de acuerdo con otros trabajos realizados anteriormente [Ruf94].

Otra alternativa a explorar es la incorporación de las características del análisis wavelets directamente en la red neuronal o etapa de clasificación con lo que se consigue una arquitectura especial de red neuronal combinada con el análisis multiresolución obtenido mediante la Transformada Wavelet. Algunos trabajos anteriores pueden tomarse como punto de partida [ZhB92], [ZWM95]. Los clasificadores híbridos (redes+árboles+lógica difusa) y el entrenamiento de las redes recurrentes con algoritmos genéticos pueden constituir otras opciones a examinar.

Como ya se dijó casi todas las aplicaciones de Wavelets existentes están orientadas al manejo de unas pocas señales (compresión, filtrado). Los inconvenientes que se presentan en la clasificación de patrones dinámicos de longitud variable como los de voz son totalmente distintos y habrá que seguir buscando soluciones más a la medida de este problema.

Un aspecto muy importante del trabajo es que se ha avanzado mucho en la definición del problema y en una serie de alternativas para encararlo. Sin duda hace falta investigar mucho más acerca de las aplicaciones de esta nueva herramienta al campo del RAH, para dar una sentencia definitiva acerca de su utilidad. El presente trabajo constituye una contribución original y novedosa en este sentido.

VII . Referencias

-[AHT93] Ahmed H. Tewfik, *"Potentials and Limitations of Wavelets in Signal Acquisition and Pocessing"*, Annals of the 1993 IEEE Engineering in Medicine & Biology 15 th Annual Conference.

-[All77] J.B. Allen, L.R. Rabiner, *"A Unified Approach to Short-Time Fourier Analysis and Synthesis"*, Proc. IEEE, Vol. 65, N°11, pp. 1558-1564, 1977.

-[ARu94] M. Argot, L. Rufiner, D. Zapata, A. Sigura, *"Una Herramienta para la Investigación Fisiológica y Clínica de la Voz y el Habla"*, Anales del XVIII Congreso Latinoamericano de Ciencias Fisiológicas, Uruguay, Abril 1994.

-[ARZ93] M. Argot, L. Rufiner, D. Zapata, A. Sigura, *"Análisis Digital de Señales de Voz Orientado a la Rehabilitación Fonoarticulatoria"*.Revista Fonoaudiológica - Tomo 39 N° 2, Octubre 1993.

-[ARZ94] M. Argot, L. Rufiner, D. Zapata, A. Sigura, *"Sistema Integrado para Fonoaudiologia"*. Anales del II Congreso Iberoamericano de Informática, Cuba, Marzo 94.

-[BCD95] J. Buckheit, S. Chen, D. Donoho, I. Johnstone, J. Scargle, *"About Wavelab"*, Stanford University and NASA-Ames Research Center, November 1995.

-[BCD95b] J. Buckheit, S. Chen, D. Donoho, I. Johnstone, J. Scargle, *"Wavelab Reference Manual"*, Stanford University and NASA-Ames Research Center, November 1995.

-[Bék60] Von Békésy G., *"Experiments in Hearing"*, McGraw-Hill, New York, 1960.

-[BeT93] J. Benedetto, A. Teolis, *"Wavelets Auditory Model and Data Compresion"*, Applied and Computational Harmonic Analysis, Vol.1 N°1, December 1993.

-[BFO84] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *"Classification and Regression Trees"*, Wadsworth. Int. 1984.

-[Bor80] Ana María Borzone de Manrique, *"Manual de Fonética Acústica"*, De. Hachette S.A., Argentina 1980.

-[Bre91] R.P. Brent, *"Fast Training Algorithms for Multilayer Neural Nets"*, IEEE Trans. Neural Networks, vol. 2, n° 3, May 1991.

-[Brow95] Dennis W. Brown, *"SPC Toolbox"*, Naval Postgraduate School, Monterey, CA 93943, dwbrown@access.digex.net, 1995.

-[CMB90] M. Cohen, H. Murveit, J. Bernstein, P. Price y M. Weintraub; *"The Decipher Speech Recognition System"* - Proc. IEEE ICASSP 1990.

-[Cod94] M.A. Cody, *"The Wavelet Packet Transform : Extending the Wavelet Transform"*, Dr. Dobb's Journal, April 1994.

-[CRZ90] J. Chaves, H. Rufiner, A. Sigura, D. Zapata, *"Reconocimiento Automático de Fonemas Basado en Redes Neuronales"*, Anales de la VII Reunión Científica de la Sociedad Argentina de Bioingeniería, 1990.

-[DaR95] C. Davidson, D. Rock, *"Wavelets and HONN: Pix-Perfect Marriage"*, AI EXPERT, Enero 1995.

-[Dau88] I. Daubechies, *"Orthonormal bases of compactly supported wavelets"*. Communications on Pure and Applied Mathematics, pp. 909-996, 1988.

-[Dau92] I. Daubechies, *"Ten Lectures on Wavelets"*, Rutgers University and AT&T Bell Laboratories, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992.

-[Dau93] I. Daubechies *"Orthonormal bases of compactly supported wavelets II. Variations on a theme"*, SIAM March 1993.

-[Dav57] Hallowel Davis; *"Biophysics and Physiology of the Inner Ear"*. Reprinted from Physiological Reviews Vol. 37 (1) January 1957.

-[DPH93] J. Deller, J. Proakis, J. Hansen, *"Discrete Time Processing of Speech Signals"*. Macmillan Publishing, NewYork, 1993.

-[DuH73] Duda, R.O., & Hart, P.E. 1973 *"Pattern Classification and Scene Analysis"*.(Q327.D83) John Wiley & Sons. ISBN 0-471-22361-1. pag. 218.

-[Elm89] J.L. Elman, *"Representation and Structure of in Connectionist Models"*, TRL Technical Report 8903, Center for Research in Language, Univ. Of California, 1989.

-[Elm90] J.L. Elman, *"Finding structure in time"*, Cognitive Science 14 (1990) 179-211.

-[ElZ88] J.L. Elman, D. Zipser, *"Learning the Hidden Structure of Speech"*, JASA 83, pp.1615-1626, 1988.

-[Fav94-2] Richard F. Favero; *"Comparison of Perceptual Scaling of Wavelet for Speech Recognition"*, SST-94.

-[Fav94-3] Richard F. Favero; *"Comparison of Mother Wavelets for Speech Recognition"*, SST-94.

-[Fav94-4] Richard F. Favero *"Compound Wavelets for Speech Recognition"*, SST-94.

-[FaG94] Richard F. Favero and Fikret Gurgen *"Using Wavelet Dyadic Grids and Neural Networks for Speech Recognition"*. ICSLP94.

-[FaK93] Richard F. Favero and Robin W. King *"Wavelet Parameterisations for Speech Recognition"*, ICSPAT93.

-[FaK94] Richard F. Favero and Robin W. King *"Wavelet Parameterisations for Speech Recognition: Variations of scale and translation parameters"*, ISSIPNN94.

-[Fav94] Richard F. Favero *"Compound Wavelets: Wavelets for Speech Recognition"*, ISTFTS94.

-[Fla89] T. Flandrin, *"Some aspects of Non Stationary Signals Processing with Emphasis on Time-Frequency and Time-Scale Methods"*, Proc. of International Conference on Wavelets, Time-Frequency Methods and Phase Space, Marselle, France, pp. 68-98, 1989.

-[Fle53] H. Fletcher, *"Speech and Hearing in Communication"*, Van Nostrand, New York, NY, 1953.

-[Fou88] J.B.J. Fourier, *"Thèorie Analytique de la Chaleur"*, Oeuvres de Fourier, Tome Premier, G. Darboux Eds., Paris, Gathlers-V. Vilars, 1888.

-[Fu74] K. S. Fu, *"Syntactic Methods in Pattern Recognition"*, Academic Press (1974).

-[FWD86] Fisher, William M., Doddington, George R., Goudie-Marshall, Kathleen M., *"The DARPA Speech Recognition Research Database : Specifications and Status"*, Proceedings of the DARPA Speech Recognition Workshop, Report N° SAIC-86/1546, February 1986, Palo Alto.

-[Gab46] D. Gabor, *"Theory of communication"*, J. Inst. Elec. Eng., vol 93, 1946

-[GeCu83] Allen Gersho and Vladimir Cuperman. *"Vector Quantization: A Pattern-Matching Technique for Speech Coding"*, IEEE Communications Magazine, pp. 15-20, December 1983.

-[GLF93] Garofolo, Lamel, Fisher, Fiscus, Pallett, Dahlgren, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentation*. National Institute of Standards and Technology, February 1993.

-[GMM95] Goddard J.C., Martinez F.M.; Martinez A.E., Cornejo J.M., Rufiner H.L., Acevedo R.C., *"Redes Neuronales y Árboles de Decisión: Un Enfoque Híbrido"*, Memorias del Simposium Internacional de Computación organizado por el Instituto Politécnico Nacional, México D.F., México, Noviembre 1995.

-[GoG95] N. González P., S.J. Garcia G., *"Uvi_Wave : Wavelets Toolbox for use with Matlab"*, Grupo de Teoría de la Señal, Universidad de Vigo, 1995.

-[GuB75] M. Guirao, A. M. Borzone, *"Identification of Argentine Spanish Vowels"*, Journal of Psycholinguistic Research , Vol 4 No 1, 1975.

-[GWS92] T. Gramβ, H. Werner Strube, *"Word Recognition with Fast Learning Neural Net"*, Applications of Neural Networks, Editado por H.G. Schuster, VCH 1992.

-[Has95] Mohamad H. Hassoun, *"Fundamentals of Artificial Neural Networks"*, The MIT Press, 1995.

-[Hel54] Helmholtz, *"Die Lehre von den Tenepfindungen als physiologische Grundlage fur die Theorie der Musik"*, Dover, Republicado 1954 (1863).

-[Kel85] J. Kelly, *"Auditory System"* en E. Kandel, J. Schwartz; Principles of Neural Science. (Elsevier, 1985)

-[KhT72] S.M. Khannam J. Tonndorf, *"Timpanic membrane vibration in cats studied by time-average holograph"*, JASA 51, 1972.

-[Koh88] T. Kohonen, *"The 'Neural' Phonetic Typewriter"*, Computer, Marzo 1988.

-[Koh92] Teuvo Kohonen *"How to make a Machine Transcribe Speech"*, Applications of Neural Networks, Editado por H.G. Schuster, VCH 1992.

-[Kol94] Eric Kolaczyk, *"Wavelet Methods for the Inversion of Certain Homogeneous Linear Operators in Presence of Noisy Data"*, Stanford Ph.D. Thesis, 1994.

-[KWT65] Kiang, Watenabe, Thomas, *"Discharge Patterns of Single Fibers in the Cat's Auditory Nerve"*, MIT Press, Cambridge, MA, 1965.

-[LaZ87] R. Lara y Zavala, "Cibernética del Cerebro", (C.E.C.S.A., 1987)

-[LDC] *Linguistic Data Consortium*, University of Pennsylvania, Philadelphia,USA, Web Site: ftp://www.cis.upenn.edu/pub/ldc_www/hpage.html, E-mail: *ldc@unagi.cis.upenn.edu.*

-[LHR90] K. Lee, A. Hauptmann, A. Rudnicky, *"The Spoken Word"*, Byte, July 1990.

-[Lip87] R. Lippmann, *"An Introduction to Computing with Neural Nets"*. IEEE ASSP Mag., Vol 4. (1987).

-[Lip89] R. Lippman, *"Review of Neural Networks for Speech Recognition"*, Neural Computation, 1 (1), 1-38.

-[Mak75] Jhon Makhoul, *"Linear Prediction: A Tutorial Review"*, Proc. IEEE, vol. 63, Nro 4, pp 561-578, April 1975.

-[Mal89] Mallat, S.G. *"A Theory of Multiresolution of Signal Decomposition: the Wavelet Representation"*, IEEE Transactions on Pattern Analisis and Machine Intelligence; Vol.11, N° 7, 1989.

-[Mar85] J. Martin, *"Cortical Neurons, the EEG and the Mechanisms of Epilepsy"*, en E. Kandel, J. Schwartz. Pinciples of Neural Science. (Elsevier, 1985)

-[MaZ93] S. Mallat, Z. Zhang, *"Matching Pursuit With Time-Frequency Dictionaries"*, IEEE Trans. in Signal Proc., December 1993.

-[McR87] J. McClelland, D. Rumelhart, *"Explorations in Parallel Distributed Processing"*, (The MIT Press, 1987).

-[Mey90] Y. Meyer, *"Ondelettes et Opérateurs"*, Tome I. Ondelettes, Herrmann de., Paris, 1990.

-[MiP69] M. Minsky, S. Papert, *"Perceptrons : An Introduction to Computational Geometry"*, MIT Press, 1969.

-[MoB95] Nelson Morgan, Hervé Bourlard, *"Continuous Speech Recognition: An Introduction to the Hybrid HMM / Connectionist Approach"*, IEEE Signal Processing Magazine, pp 25-42, vol. 12, N° 3, May 1995.

-[Moz92] M.C. Mozer, *"Induction of multiscale temporal structure"*, Advances in Neural Information Processing Systems 4, Moody, Hanson, Lippmann, Eds., San Mateo, CA : Morgan Kaufmann, 1992.

-[MST94] D. Michie, D.J. Spiegelhalter, C.C. Taylor, *"Machine Learning, Neural and Statistical Classification"*, Ellis Horwood 1994.

-[NaR95] Abinash Nayak, Rob J. Roy, *"Neural Networks for Predicting Depth of Anesthesia from Auditory Evoked Potentials: A Comparison of the Wavelet Transform with Autoregressive Modeling and Power Spectrum Feature Extraction Methods"*, Annals of the 1995 IEEE Engineering in Medicine & Biology 17 th Annual Conference.

-[Neu2.3] NeuroShell 2 Release 3.0, Ward Systems Group, Inc. 1996.

-[NgW89] D. Nguyen, B. Widrow, *"The truck backer-upper : An example of self-learning in neural networks"*, Proc. of the International Joint Conference on Neural Networks, Vol. II, pp.357-362, 1989.

-[Now88] S.J. Nowlan, *"Gain variation in recurrent error propagation networks"*, Complex Systems, 2, pp.305-320, 1988.

-[OGI] *Oregon Graduate Institute*, Center for Spoken Language Understanding, Portland, USA, Web Site: http://www.cse.ogi.edu/CSLU/corpora/isolet2.html.

-[Ope70] Alan V. Oppenheim, *"Speech Spectrograms using the fast Fourier transform"*, IEEE Spectrum, Agosto 1970.

-[PeB52] Peterson, Barney, *"Control methods used in a study of the vowels"*, JASA 24, 175-184, 1952.

-[Por80] M.R. Portnoff, *"Time-Frecuency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis"*, IEEE Trans. on Acoust., Speech and Signal Proc., Vol. 28, pp. 55-69, Feb. 1980.

-[Qui93] J.R. Quinlan, *"C4.4"*, Morgan Kaufmann 1993.

-[RaG75] L. Rabiner, B. Gold, *"Theory and Application of Digital Signal Processing"*, (Prentice-Hall. 1975)

-[RaS87] L. Rabiner, R. Schafer, *"Digital Processing of Speech Signals"*, (Prentice-Hall, 1987)

-[RGT94] L. Rufiner,L. Gamero, M.E. Torres, D. Zapata, A. Sigura, *"Comparison Between Wavelets and Fourier Analysis as Speech Recognition Preprocessing Techniques"*, Annals of the World Congress on Medical Physics and Biomedical Engineering, Part 1 pp. 278, Agosto 1994.

-[RHPF] T. Robinson, J. Holdsworth, R. Patterson, F.Fallside, *"A comparison of preprocessors for the cambridge recurrent error propagation network speech recognition system"*, Cambridge University Engineering Department.

-[RHW86] D. Rumelhart, G. Hinton, R. Williams. *"Learning Internal Representations by Error Propagation"*, Nature, vol 323, pp 533-536, Oct 1986.

-[RiV91] Olivier Rioul, Martin Vetterli *"Wavelets and Signal Processing"*, IEEE Magazine on Signal Processing, pp. 14-38, Octubre 1991.

-[Roc87] L. Rocha, *"Procesamiento de voz"*, I Escola Brasileiro-Argentina de Informática, (Kapelusz, 1987).

-[RSZ93] Rufiner, Sigura, Zapata, *"Interfaz oral para computadoras personales orientada al discurso continuo"*, Revista Telegráfica Electrónica, N° 938, 602-609, Editorial Arbó, Enero 1993.

-[RTZ91] L. Rufiner, C. Tabernig, D. Zapata, *"Desarrollo de un Analizador de secuencias fonéticas aplicable al Reconocimiento automático del habla"*, Anales de la IV Reunión de Trabajo en Procesamiento de la Información y Control (RPIC '91) - (CONEA), Argentina - Noviembre 1991.

-[Ruf94] Rufiner H. Leonardo, *"Modelización Biológica, Redes Neuronales y HMM's aplicados al Reconocimiento Automático del Habla"*, Informe de Avance Beca de Investigación CONICET, Argentina, 1994.

-[Rug92] Ruggero M.A., *"Responses to sound of basilar membrane of mammalian cochlea"*, Current Opinion Neurobiol. 2, 1992.

-[RuZ91] L. Rufiner, D. Zapata, *"Spanish Speech Recognition using Neural Networks and Parsing Techniques"*, Annals of the IV International Symposium on Biomedical Engineering, España, (LIB/UPV/AEBI/IFBME), Septiembre 1991.

-[RuZ92] H. Rufiner, D. Zapata, *"Desarrollo de un Sistema de Reconocimiento Automático del Discurso Continuo, Independiente del Hablante y con Vocabulario Ampliable"*. Tesis de grado de la Carrera de Biongeniería de la UNER, Argentina, Mayo de 1992.

-[RZa91] L. Rufiner, D. Zapata, *"Redes neuronales aplicadas al reconocimiento del discurso continuo"*. Anales de la IV Reunión de Trabajo en Procesamiento de la Información y Control (RPIC '91) - (CONEA), Argentina - Noviembre 1991.

-[RZa92] L. Rufiner, D. Zapata, A. Sigura, *"Simulación, Entrenamiento y Evaluación de Redes Neuronales Anteroalimentadas en Computadoras Personales"*, Anales del Primer Congreso Conjunto de Bioingeniería y Física Médica, VIII Congreso Argentino de Bioingeniería y III Workshop de Física Médica, (SABI /SAFIM), Octubre de 1992.

-[RZS92] L. Rufiner, D. Zapata, A. Sigura, *"Sistema de Adquisición, Procesamiento y Análisis de Señales de Voz"*, Anales del Primer Congreso Conjunto de Bioingeniería y Física Médica, VIII Congreso Argentino de Bioingeniería y III Workshop de Física Médica, (SABI/SAFIM), Octubre de 1992.

-[RZS93] L. Rufiner, D. Zapata, A. Sigura, *"Interfaz oral para computadoras personales orientada al discurso continuo"*, Revista Telegráfica Electrónica, N° 938, 602-609, Editorial Arbó, Enero 1993.

-[SaC94] N. Saito, R.R. Coifman, *"Local Discriminant Bases"*, Mathematical Imaging : Wavelet Applications in Signal and Image Processing II, A.F. Laine, M.A.Unser, Editors, Proc. SPIE Vol. 2303, 1994.

-[SaM91] A. Sankar, R. Mammone, "Neural Tree Networks", Neural Networks, Mammone and Zeevi Editors, Academic Press, 1991.

-[ScR75] R. Schafer, L. Rabiner, *"Digital Representations of Speech Signals"*, IEEE Proc. Vol 63 No 4 (1975)

-[Sch75] Manfred R. Schroeder; *"Models of Hearing"*. Proceedings of the IEEE Vol. 63 (9) September, 1975.

-[Sch84] Manfred R. Schafer, *"Linear Predicction, Entopy and Signal Analysis"*, IEEE, ASSP Magazine, pp. 3-11, July 1984.

-[Sej86] T. Sejnowski, *"Open Questions About Computation in Cerebral Cortex"*, Parallel Distributed Processing, Vol 2 (The MIT Press, 1986)

-[Sel85] P. Seligman, *"Speech-Processing Strategies and their Implementation"* , 1985

-[Set90] I. K. Sethi, *"Entropy nets: from decision trees to neural nets"* Proceedings of the IEEE, Vol.78, pp. 1605-1613, 1990.

-[Set91] I. K. Sethi, *"Decision tree performance enhancement using an artificial neural network implementation"*, Artificial Neural Networks and Statistical Pattern Recognition Old and New Connections, Elsevier Science Publishers B.V., pp 71-88, 1991.

-[SGL92] J. Schuchhardt, J.C. Gruel, N. Lüthje, L. Molgedey. G. Radons y H.G.Schuster, *"Neural Networks for the Classification of Sound Patterns"*, Applications of Neural Networks, Editado por H.G. Schuster, VCH 1992.

-[Sim87] G. Simons, *"Introducción a la Inteligencia Artificial"* (1987).

-[SMD83] A. Stirnemann ; G. S. Moschitz ; N. Dillier, *"A Network model of the Middle Ear"*. Swiss Federal Institute of Technology, Zurich, Switzerland. September 1983.

-[SoC95] M.N. Souza, L.P. Caloba, *"An Analytical Auditory Wavelet"*, Annals of the 13th Brazilian Telecommunications Symposium, Águas de Sindorá, 1995.

-[SoC96] M.N. Souza, L.P. Caloba, *"A Comparison between Fourier and Biological Based Time-Frequency Distributions, Applied to the Speech Signals"*, Annals of the 39th Midwest Symposium on Circuits and Systems, Ames, 1996.

-[Som86] G. Somjen, *"Neurofisiología"*, Editorial Médica Panamericana, Buenos Aires 1986.

-[SoV93] A.K. Soman, P.P. Vaidyanathan, *"On orthonormal wavelets and paraunitary filter banks"*, IEEE Trans, Signal Proc., vol. SP-41, pp.1170-1183, March 1993.

-[Str93] R.S. Strichartz, *"How to make wavelets"*, American Mathematical Monthly, June-July, pp.539-556, 1993.

-[SWS90] Secker-Walker, Searle, *"Time-domain analysis of auditory-nerve-fiber firing rates"*, JASA 88, 1990.

-[TaH87] D.W. Tank, J.J. Hopfield, *"Concentrating information in time : Analog Neural Networks with Applications to Speech Recognition Problems"*, IEEE First International Conference on Neural Networks, 1987.

-[Tak95] Haruhisa Takahashi, *"Voice Recognition using Recurrent Neural Networks"*., Dep. of Comunications and Systems, University of Electrocomunications.

-[Tas95] C. Taswell, *"Speech Compression with Cosine and Wavelet Packet Near-Best Bases"*, Stanford University, 1995.

-[TNN94] *IEEE Trans. on Neural Networks* : Special Issue on Recurrent Networks, Vol.5, N°2, 1994.

-[Tra86] Bryan J. Travis, *"A Layered Neural Network Model Applied to the Auditory System"*, en Neural Networks for Computing, AIP Conference Proceedings 151 (1986).

-[UAM91] M. Unser, A Aldroubi., E. Murray, *"Fast B-Spline Transform for Continuous Image Representation and Interpolation"*, IEEE Trans on Pattern Anal. and Machine Intell., Vol 13, No 3, pp 277-285, March 1991.

-[UAM93] M. Unser, A Aldroubi., E. Murray, *"A Family of Polinomial Spline Wavelet Transform"*, Signal Processing 30, 141-162, Elsevier, 1993

-[Vai92] P.P. Vaidyanathan, *"Multirate Systems and Filter Banks"*, Prentice-Hall, pp. 532-535. 1992.

-[Ver74] S. Hirzel Verlag-Stuhgart *"Two Formant Models, Pitch, and Vowel Perception"*. Acustica Vol. 31 (6) 1974.

-[Vie80] Max Viergever, *"Mechanics of the inner ear"*. Edic. 1980.

-[WaH89] A. Waibel, J. Hampshire, *"Building Blocks for Speech"*, BYTE, Agosto 1989.

-[WeW92] E. Wesfreid, M.V. Wickerhauser, *"Adapted Local Trigonometric Transforms and Speech Processing"*, Université Paris IX-Dauphine, 1992.

-[WHH89] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang; *"Phoneme Recognition Using Time-Delay Neural Networks"*. IEEE Trans. ASSP Vol. 37, No 3 (1989).

-[Whi90] G.M. White, *"Natural Language Understanding and Speech Recognition"*, Comunications of the ACM, Vol 33, N°8, 1990.

-[Wic91] M.D. Wickerhauser, *"Acoustic Signal Compression with Wave Packets"*, Yale University, 1991.

-[Wic91b] M.D. Wickerhauser, *"Lectures on Wavelet Packet Algorithms"*, Washington University, Nov. 1991.

-[Wil93] W.H. Wilson, *"A comparison of Architectural Alternatives for Recurrent Networks"*, Proceedings of the Fourth Australian Conference on Neural Networks, pp.189-192, 1993.

-[Wil95] W.H. Wilson, *"Stability of Learning in Classes of Recurrent and Feedforward Networks"*, Proceedings of the Sixth Australian Conference on Neural Networks, pp.142-145, 1995.

-[WiL90] B. Widrow, M. Lehr, *"30 Years of Adaptive Neural Networks : Perceptron, Madaline and Backpropagation"*, Proceedings of IEEE. Vol. 78, No. 9, pp.1415-1442, Sept. 1990.

-[WSS89] Waibel A. H. Sawai, Shikano, *"Consonant Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks"*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Glasgow, Scotland, vol. 1, pp. 112-115, 1989.

-[ZhB92] Qinghua Zhang, Albert Benveniste, *"Wavelets Networks"*, IEEE Transactions on Neural Networks, pp. 889, vol. 3, N° 6, Noviembre 1992.

-[ZWM95] Jun Zhang, Gilbert G. Walter, Yubo Miao, and Wan Ngai Eayne Lee, *"Wavelet Neural Networks for Function Learning"*, IEEE Transactions on Signal Processing, vol 43, N° 6, Junio 1995.

Nota : Varios de los artículos que aparecen en estas referencias han sido obtenidos a través de Internet. En los casos donde no aparece cita a la fuente original se preser... el grupo que desarrolló el trabajo y la universidad donde se realizó, por ser estos los datos más importantes para ubicar el documento.