UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO

FACULTAD DE INGENIERÍA ELÉCTRICA

TESIS

SISTEMA AUTOMÁTICO DE RECONOCIMIENTO DE VOZ

Que para obtener el grado de
MAESTRO EN INGENIERÍA ELÉCTRICA

Presenta
Ismael Chávez Álvarez

Dr. José Antonio Camarena Ibarrola
Director de Tesis

Mayo 2009

Resumen

El reconocimiento de voz constituye uno de los campos en que la computación técnica mantiene un desarrollo constante con el propósito de explorar soluciones cada vez más útiles al problema de la comunicación hombre-máquina. La combinación de las diversas áreas del conocimiento que esta dificultad abarca, expande el horizonte de posibilidades a medida que se explora cada uno de los sub-problemas implicados, y con ello se enriquece la calidad de las propuestas de resolución del problema. Este proyecto efectúa el reconocimiento automático de palabras aisladas, enfatizando en el componente que se especializa en la caracterización de las señales acústicas para que estas puedan ser manipuladas por otros módulos que integran un sistema de reconocimiento de voz completo.

Se describe el uso de los coeficientes *Wavelet* obtenidos al aplicar la *Transformada Wavelet Discreta* a una señal de voz, como parte de un sistema automático de clasificación de palabras. La Transformada Wavelet Discreta que se emplea en esta investigación incorpora como Wavelets madre a la función de *Haar* y las funciones miembros de la familia de Wavelets de *Daubechies* de ordenes 2 a 10, e implementa los algoritmos piramidales propuestos por Stéphane Mallat. Se detallan las distintas fases de la construcción de un sistema simple de reconocimiento de palabras pronunciadas aisladamente, entre las que destaca la etapa de extracción de características de la señal de la voz humana, etapa en la cual se implanta la transformación antes mencionada. Se presenta la evaluación del rendimiento del sistema haciendo uso de la *Teoría de Detección de Señales*, que tiene como su expresión gráfica más conocida a las curvas ROC (*Receiver Operating Characteristic*) como un medio para la elección de un modelo posiblemente óptimo y descartar otros posiblemente subóptimos.

De las pruebas realizadas se han obtenido resultados satisfactorios, al corroborar que es pertinente el empleo de la herramienta de análisis propuesta, los Wavelets, para resolver el problema consistente en encontrar las características en la voz humana. Se construyeron diversos clasificadores que difieren unos de otros en el componente clave utilizado para la etapa de caracterización, y se ha distinguido cuáles de ellos son los que reportan mejor rendimiento. La evidencia encontrada señala en la dirección de que un análisis multiresolución simple de la señal acústica permite la representación alterna de la señal original mediante datos los cuales, no obstante su presencia en número reducido, transportan to-

davía información significativa de la señal, y esto puede emplearse provechosamente para los fines de clasificación y de reconocimiento automáticos.

Abstract

Speech recognition is a field in which technical computing keeps steady development aimed to survey even better solutions to the man-machine communication problem than the ones available to date. The mixture of different knowledge fields involved in this problem widens the horizon of possibilites as each one of the implied sub-problems is revised, and enriches the quality of the proposals to solve the problem. This project performs automatic isolated word recognition paying much attention in developing the component specialized on discovering the features contained in acoustic signals, as for them to be manipulated by later stages in a complete voice recognition system.

This document describes the use of *Wavelet* coefficients obtained by applying the *Discrete Wavelet Transform* on a voice signal, as part of an automatic word classifying system. The Discrete Wavelet Transform used in this research embodies as mother Wavelets the *Haar* function and members of the *Daubechies* Wavelet functions orders 2 thru 10, and implements the piramidal algorithms proposed by Stéphane Mallat. All the different stages implied in the construction of a simple isolated pronounced words recognition system are shown, among which, characteristics extraction from human voice signal becomes paramount, being this stage the one that implements the aforementioned transform. The performance of the system is shown with the use of *Signal Detection Theory* which is widely represented graphically using ROC (*Receiver Operating Characteristic*) curves, as a means to choose between a possibly optimal model and discard possibly suboptimal ones.

We are pleased with the results from the tests performed, as we notice that the choice of Wavelets for an analysis tool is well suited to solve the feature extraction problem in human voice. Several different classifiers were built changing the key component used to perform the feature extraction stage in the system, and the best ones were identified. Collected evidence indicates that a simple multiresolution analysis on acoustic signals enables an alternate representation of the original signal thru data that, despite its presence in short number, still carries important information from the signal, and this can be used advantageously in the aim of automatic classification and recognition.

Contenido

Resumen III
Abstract V
Contenido VII
Lista de Figuras XI
Lista de Tablas XIII
Lista de Algoritmos XV
Lista de Símbolos XVII

1. Introducción 1
   1.1. Planteamiento del Problema 1
      1.1.1. El reconocimiento de voz 1
   1.2. Antecedentes 2
      1.2.1. Introducción 2
      1.2.2. Los problemas en el reconocimiento automático de la voz 4
      1.2.3. Arquitectura de un sistema de reconocimiento de palabras aisladas . 6
      1.2.4. Tipos de sistemas de reconocimiento de voz 8
   1.3. Objetivos de la Tesis 12
      1.3.1. Objetivo general 12
      1.3.2. Objetivos particulares 12
   1.4. Descripción de Capítulos 13
      1.4.1. Organización de este documento 13

2. El procesamiento digital de la señal de voz 15
   2.1. Segmentación de la señal de voz 15
      2.1.1. Energía de tiempo corto 16
      2.1.2. Magnitud promedio de tiempo corto 17
      2.1.3. Régimen de cruces por cero de tiempo corto 17
   2.2. Transformadas de Fourier y de Fourier de tiempo corto 18
      2.2.1. Transformada de Fourier 18
      2.2.2. Desventajas en el uso de la DFT para el análisis de la señal de voz . 19
      2.2.3. La Transformada de Fourier Discreta de Tiempo Corto 22
   2.3. Wavelets 26
      2.3.1. Análisis con Wavelets 26
      2.3.2. Transformada Wavelet Continua 28

2.3.2.1. Escala y análisis multiresolución . . . 30
2.3.2.2. Traslación . . . 31
2.3.3. Transformada Wavelet Discreta . . . 33
2.3.3.1. Aproximaciones y detalles . . . 35
2.3.3.2. Descomposición multinivel . . . 38
2.3.4. El Wavelet de Haar . . . 39
2.3.5. Los Wavelets de Daubechies . . . 41
2.4. Implantación de un algoritmo de DWT . . . 41
2.4.1. DWT basada en el Wavelet de Haar . . . 41
2.4.2. DWT basada en el Wavelet de Daubechies de orden 2 . . . 42
2.5. Caracterización de la señal de voz con Wavelets . . . 43

3. Similitud y clasificación de patrones 47
3.1. Similitud entre dos patrones . . . 48
3.2. Distancia entre dos vectores de características de la señal de voz . . . 50
3.2.1. Distancias L_p . . . 51
3.2.1.1. Distancia L_1 . . . 51
3.2.1.2. Distancia euclidiana o L_2 . . . 52
3.3. Alineación temporal . . . 53
3.3.1. Doblado Dinámico en Tiempo . . . 54
3.4. Clasificación de patrones . . . 58
3.5. El vecino más cercano . . . 58
3.6. Los k-vecinos más cercanos . . . 59
3.7. Diccionario . . . 60

4. Sistema desarrollado 63
4.1. Captura de la señal . . . 64
4.2. Segmentación de la señal . . . 65
4.3. Extracción de características . . . 68
4.4. Medición de distancia . . . 71
4.5. Clasificación . . . 72

5. Resultados 75
5.1. Gráficas ROC (Receiver Operation Characteristic) . . . 75
5.1.1. Rendimiento de un clasificador . . . 75
5.1.2. Espacio ROC . . . 78
5.2. Experimentos realizados . . . 81
5.3. Caracterización de palabras mediante coeficientes de aproximación obtenidos con transformada basada en el wavelet de Haar (db01). . . . 82
5.4. Caracterización de palabras mediante coeficientes de aproximación obtenidos con transformada basada en el wavelet de Daubechies (db02 a db10). . . . 87
5.5. Características de los clasificadores a partir de la información revelada por las curvas ROC. . . . 87

6. Conclusiones 99
6.1. Conclusiones Generales 99
6.2. Conclusiones Específicas 100
6.3. Trabajos Futuros 101

A. Construcción de un Wavelet 103
A.1. Primera iteración 103
A.2. Segunda iteración 104
A.3. Tercera iteración 105
A.4. Cuarta iteración 106

Referencias 111

Glosario 115

Lista de Figuras

1.1. Estructura típica de un sistema de reconocimiento de voz. 7

2.1. Señales sinusoidales puras. Espectros de magnitudes de las DFT correspondientes a las señales sinusoidales puras. 20
2.2. Señales estacionaria y no-estacionarias y sus correspondientes espectros de magnitudes por DFT. 20
2.3. Tipos de ventanas empleadas para seccionar señales. 23
2.4. El análisis de Fourier de Tiempo Corto ofrece una resolución fija tanto en tiempo como en frecuencia. 25
2.5. El análisis con Wavelets ofrece una resolución variable tanto en frecuencia como en tiempo. 27
2.6. Funciones base del análisis de Fourier y del análisis con Wavelets. 28
2.7. Coeficientes Wavelet para el análisis en escala continua para una elocución de la palabra *"uno"*. 29
2.8. Wavelet madre de Daubechies de orden 2 en tres escalas distintas. 32
2.9. Traslación del wavelet madre. 33
2.10. Derivación de aproximaciones A y detalles D mediante el proceso de filtrado de la señal S. 36
2.11. Filtrado de una señal S sin submuestreo para la obtención de A y D. 37
2.12. Obtención de los Coeficientes Wavelet de Aproximación cA y Detalle cD mediante el filtrado con submuestreo de una señal S. 37
2.13. Descomposición multinivel de una señal S. 38
2.14. Descomposición multinivel de una elocución de la palabra *"uno"*. 39
2.15. Wavelet de Haar. 40

3.1. Escalagramas de la elocución de dos palabras distintas: *"uno"* y *"cero"*. . . 48
3.2. Escalagramas de dos elocuciones de la misma palabra: *"uno"*. 49
3.3. Interpretación geométrica de algunas de las distancias L_p. 52
3.4. Matriz de alineación de patrones con Doblado Dinámico en Tiempo. 55
3.5. Tres movimientos posibles de un punto al siguiente en el trayecto de alineación temporal empleando restricciones locales simétricas de primer orden. . 56
3.6. Restricciones locales usadas para DTW. 57

4.1. Señal obtenida de la etapa de captura del sistema. 65

4.2. Segmentación de la señal con la estrategia del régimen de cruces por cero. . 66

5.1. Matriz de confusión. 76
5.2. Distintos clasificadores en el espacio ROC. 79
5.3. Distancias entre los patrones de 10 instancias de pronunciación de 10 clases distintas palabras y el patrón de la palabra *"uno"*. 83
5.4. Distancias promedio entre los patrones de 10 instancias de pronunciación de 10 clases distintas palabras y el patrón de la palabra *"uno"*. 84
5.5. Curvas ROC para cinco clasificadores basados en el Wavelet de Haar. . . . 87
5.6. Curvas ROC para cinco clasificadores basados en el Wavelet db02. 88
5.7. Curvas ROC para cinco clasificadores basados en el Wavelet db03. 88
5.8. Curvas ROC para cinco clasificadores basados en el Wavelet db04. 89
5.9. Curvas ROC para cinco clasificadores basados en el Wavelet db05. 89
5.10. Curvas ROC para cinco clasificadores basados en el Wavelet db06. 90
5.11. Curvas ROC para cinco clasificadores basados en el Wavelet db07. 90
5.12. Curvas ROC para cinco clasificadores basados en el Wavelet db08. 91
5.13. Curvas ROC para cinco clasificadores basados en el Wavelet db09. 91
5.14. Curvas ROC para cinco clasificadores basados en el Wavelet db10. 92
5.15. Curvas ROC para diez clasificadores basados en los wavelets de Haar y de Daubechies de ordenes 2 a 10, todos ellos en escala 5. 94
5.16. Curvas ROC para tres clasificadores basados en el Wavelet de Daubechies de escala 4 y además un clasificador basado en el Wavelet de Daubechies de escala 3. 95
5.17. Curvas ROC para diez clasificadores basados en los wavelets de Haar y de Daubechies de ordenes 2 a 10, todos ellos en escala 4. 96

A.1. Primera iteración para la generación de la forma de onda del Wavelet db02. 109
A.2. Segunda iteración para la generación de la forma de onda del Wavelet db02. 109
A.3. Tercera iteración para la generación de la forma de onda del Wavelet db02. 110
A.4. Cuarta iteración para la generación de la forma de onda del Wavelet db02. . 110

Lista de Tablas

2.1. Correspondencia escala de wavelets-contenido de frecuencia de señales. . . . 32

3.1. Entrada de diccionario correspondiente a la palabra *"cero"* caracterizada con Wavelet de Daubechies de orden 2 y nivel 5. 61

4.1. Coeficientes de los filtros pasabajas de descomposición de los Waveles de Daubechies de ordenes 1 a 4. 69
4.2. Coeficientes de los filtros pasabajas de descomposición de los Waveles de Daubechies de ordenes 5 a 7. 69
4.3. Coeficientes de los filtros pasabajas de descomposición de los Waveles de Daubechies de ordenes 8 a 10. 70

5.1. 28 vecinos más cercanos a la elocución de la palabra *"uno"* y 99 otras elocuciones de todas las palabras correspondientes a los dígitos, *"uno"* inclusive. 85

Lista de Algoritmos

1. DWT usando el wavelet de Haar 42
2. DWT usando el wavelet de Daubechies de orden 2 44
3. Doblado Dinámico en Tiempo 72

Lista de Símbolos

E	Energía de tiempo corto.
M	Magnitud promedio de tiempo corto.
Z	Régimen de cruces por cero de tiempo corto.
$X(f)$	Función del dominio de la frecuencia.
$x(t)$	Función del dominio del tiempo.
Ψ	Función Wavelet.
t	Tiempo.
C	Coeficientes Wavelet.
a	Factor de escalamiento de un Wavelet.
Φ	Conjunto de funciones base formadas por el escalamiento y la traslación de un Wavelet madre.
s	Dilatación del Wavelet madre.
l	Traslación del Wavelet madre.
W	Función de escalamiento del Wavelet madre para transformación en tiempo discreto.
c_k	Coeficientes de definición del Wavelet (coeficientes del filtro).
S	Señal bajo estudio.
A	Aproximación a la señal.
D	Detalles de la señal.
cA	Coeficientes de aproximación de la transformada con Wavelets.
cD	Coeficientes de detalle de la transformada con Wavelets.
$L_0, L_1, L_2, L_3, ...$	Coeficientes de definición del Wavelet para el filtro pasabajas L.
d_{L_p}	Clases de distancias entre vectores.
N	Dimensión de un vector.
d_{L_1}	Distancia en espacio L_1.
d_{L_2}	Distancia en espacio L_2.
$D(i,j)$	Distancia del punto i al punto j.
$\{p_1, p_2, ...\}$	Conjunto de puntos en un espacio.

$\{p_a, p_b, ...\}$	Conjunto de puntos en un espacio.
PF	Positivos Falsos.
PV	Positivos Verdaderos.
NV	Negativos Verdaderos.
PT	Positivos en Total.
NT	Negativos en Total.
$db01$	Wavelet de Haar (Daubechies de orden 1).
$db02, ..., db10$	Wavelets de Daubechies de ordenes 2 a 10.

Capítulo 1

Introducción

El propósito de este capítulo es el de describir brevemente los antecedentes y la situación actual del complejo problema del reconocimiento de la voz humana. Se menciona un conjunto de condiciones que hacen árdua la tarea del reconocimiento de voz. Del mismo modo, se comenta la estructura elemental de todo sistema de este tipo. Una vez hechas estas consideraciones, se describe el alcance que se espera en el desempeño del prototipo computacional que se construyó como un sistema autónomo de reconocimiento de palabras aisladas empleando el lenguaje de programación de computadoras Java, y cuyos detalles se presentan en el capítulo 4.

1.1. Planteamiento del Problema

1.1.1. El reconocimiento de voz

La comunicación oral entre personas constituye un mecanismo natural, económico y eficaz del que dispone nuestra especie para la trasmisión de las ideas. Es notable el resultado evolutivo que hoy nos permite extraer un mensaje complejo que se halla contenido en una onda sonora, caracterizada por causar variaciones en la presión del aire en la vecindad de nuestros oídos.

Uno de los grandes retos que la tecnología de nuestros días lleva a cuestas, es el de extender el proceso de comunicación oral para que no solamente sea posible entre humanos,

sino que contemos con la oportunidad de transmitir mensajes a aparatos, máquinas y herramientas. Este es sin duda el proyecto más ambicioso que mantienen en curso los expertos en tecnología de procesamiento del habla, cuya meta es la de "conversar" con una máquina de manera continua con toda naturalidad, con el consiguiente reconocimiento y entendimiento de las ideas en ambos sentidos.

No obstante, existen otras tareas relativamente menos ambiciosas, que también son preocupación de los equipos de desarrollo de sistemas relacionados con el habla, tales como la conversión básica texto-voz y voz-texto. Es en este contexto donde se ubica el problema del reconocimiento de los símbolos del lenguaje, los cuales están asociados con las características de la onda sonora que transporta el mensaje. Con una solución satisfactoria de esta dificultad, se tiene entonces la posibilidad de emitir comandos orales simples, que permitan una interacción hombre-máquina en los planos de naturalidad y confort a los que las personas estamos habituadas cuando nos comunicamos.

1.2. Antecedentes

1.2.1. Introducción

La aparición y la consecuente disponibilidad generalizada de computadoras digitales cada vez más poderosas, ha hecho posible el surgimiento de estudios de modelado cuantitativo que están basados en las propiedades fundamentales del habla humana, y cuyos resultados son cada vez más halagadores.

Los primeros proyectos que abordaron el estudio de este tipo de problema se presentaron a finales de la década de los 1960's. Del resultado del proyecto ARPA-SUR (Advanced Research Projects Agency-*Speech Understanding Research*) [Lea79] desarrollado bajo el auspicio de la Agencia de Proyectos de Investigación Avanzados del Departamento de la Defensa de los Estados Unidos, se implementaron los sistemas DRAGON, HEARSAY [Reddy76] y HARPY [Reddy89]. IBM (International Business Machines) desarrolló LASER en 1980. Otros proyectos relevantes son los desarrollados por DARPA bajo supervisión del NIST (National Institute of Standards and technology-Instituto Nacional de Estándares y Tecnología); y SAM (Speech Assessment Methodology-Metodología de Evaluación del

Habla).

Actualmente, los grupos de investigación en la Universidad Carnegie Mellon CMU (Carnegie Mellon University) desarrollan proyectos tales como LISTEN (*Literacy Innovation that Speech Technology ENables*) patrocinado por la Fundación Nacional de la Ciencia de los Estados Unidos NSF (National Science Foundation), que consiste en un tutor automático de lectura el cual despliega una historia en la pantalla de una computadora, escucha a un niño leerla en voz alta, y brinda ayuda cuando se hace necesario [CMULPG07]. Otro proyecto que se se encuentra en curso en CMU bajo el patrocinio de DARPA es SPHINX-3 y 4 [CMURSRG07], cuyos orígenes se remontan a la década de los 1980's, y que en palabras de sus autores *"es uno de los mejores y más versátiles sistemas de reconocimiento en el mundo en la actualidad"*. El proyecto SPHINX fue el primer sistema que permitió demostrar la factibilidad de reconocimiento de discurso continuo independiente del usuario y de vocabulario amplio, posibilidad de lo cual se encontraba en debate en la época (1986). SPHINX utilizó Modelos Ocultos de Markov y parámetros derivados por Codificación Lineal Predictiva [Lee90]. SPHINX-II es un reconocedor veloz orientado al rendimiento, enfocado en el reconocimiento de tiempo-real que fuese propicio para las aplicaciones de lenguaje hablado. SPHINX-II cuenta entre sus características el intercambio dinámico de modelo del lenguaje [Huang92]. SPHINX-3 utiliza una representación semi-continua para el modelado acústico, y aunque inicialmente se le usó para el reconocimiento de precisión de no-tiempo-real, los avances recientes en equipo y algoritmos de cómputo lo han convertido en un sistema cercano-a-tiempo-real que se mantiene en desarrollo activo [CMURSRG07]. SPHINX-4 constituye una reescritura completa del motor de SPHINX con el propósito de proporcionar un marco de trabajo más flexible para la investigación en reconocimiento del habla y se encuentra en desarrollo activo [CMURSRG07]. PocketSphinx es una versión de SPHINX que se puede utilizar en sistemas empotrados (por ejemplo, los que están basados en procesadores ARM) y que se encuentra en desarrollo activo para proveer de características tales como las de aritmética de punto fijo y algoritmos eficientes para el cálculo de Modelos de Mezcla de Gaussianas [CMURSRG07]. En el Centro para la Investigación del Lenguaje Hablado de la Universidad de Colorado en Boulder se desarrollan PHOENIX [Ward91] y el proyecto CU COMMUNICATOR, que tiene por misión *"el desarrollar una radicalmente novedosa capa-*

cidad en la disciplina de tecnología de la información que permita a las personas hablar con computadoras", ello con el propósito de crear, accesar y manejar información, y de resolver problemas de manera completa [Ward99, Pellom00, UCCSLR07]. Parte de estos grupos se encargan del área de *Reconocimiento Robusto*, cuya preocupación fundamental consiste en la elaboración de sistemas de reconocimiento del discurso en ambientes adversos.

El proyecto JANUS es un sistema de traducción de lenguaje oral que opera en una manera similar a un intérprete humano [CMUISL07, Waibel96].

1.2.2. Los problemas en el reconocimiento automático de la voz

Existe una gama amplia de dificultades que debe tratar de mantenerse suficientemente bajo control cuando se ataca la tarea de reconocer la voz humana con un mecanismo automático artificial. Entre estas se puede mencionar:

- **La limitación sensorial que tienen las máquinas.**

  El problema de la *percepción* del entorno. La interpretación por parte de las personas de las ideas contenidas en sonidos que se transmiten por el aire requiere de la activación y funcionamiento de diversos y complejos órganos en el cuerpo. El intrincado funcionamiento de los sentidos en los seres humanos, funcionamiento que se intenta replicar por los dispositivos electrónicos, no se ha comprendido con plenitud todavía. Es notable el proceso de pensamiento necesario para la comprensión del significado de los impulsos eléctricos que procesa el cerebro en la comunicación oral. En efecto, el sentido del oído en el ser humano hace mucho más que solamente escuchar conversaciones.

- **La extrema complejidad de las ondas sonoras que viajan en el aire.**

  El problema del *modelado* del entorno. Lo que el oído percibe no son solamente las palabras que un locutor pronuncia, sino que se percibe la composición de la voz con una gran cantidad de emisiones de otras fuentes diferentes, o bien, información que proviene de la misma fuente pero que no constituye parte del mensaje que se desea emitir. A partir de esta masa confusa de estímulos sensoriales es que

se tendrá que discernir entre los componentes adecuados para la construcción de un modelo conveniente y los datos redundantes o inconvenientes.

- **La naturaleza continua de la voz humana.**

  El problema de la *segmentación* de los elementos del lenguaje. Aún cuando pueda aislarse absolutamente la fuente de emisión de un mensaje oral, persiste el problema de identificar sus componentes mínimos con los cuales se pueda efectuar el reconocimiento eficientemente. Aún cuando sabemos que gramaticalmente los símbolos del lenguaje consisten de letras, las cuales unimos para formar sílabas, y que con estas últimas formamos palabras las cuales a su vez están separadas entre sí por espacios, al escuchar, lo que normalmente llega a nuestros oídos es un torrente ininterrumpido de sonidos sin separaciones entre ellos.

- **La naturaleza contextual de los sonidos en la voz humana.**

  El problema de la *coarticulación* de los sonidos. Dos sonidos que se producen al pronunciar dos palabras diferentes, aún cuando correspondan a un único símbolo del lenguaje, pueden poseer características completamente distintas dependiendo de cuáles son los símbolos que le anteceden y cuáles los que le suceden, dicho de otra manera, los vecinos de un símbolo pueden afectar como suena este cuando se pronuncian palabras diferentes.

- **La unicidad de la voz de cada persona.**

  El problema de la *dependencia* del usuario. El sonido que una persona produce al pronunciar una palabra, no solamente es el vehículo que aprovecha la semántica del lenguaje para que se transmita una idea, sino que también sirve como medio de transporte de algún "sello personal" que nos hace capaces de distinguir entre fuentes distintas. Esto es una especie de "huella digital oral" que hace que dos mensajes exactamente iguales se transporten mediante dos ondas sonoras diferentes. Existe una gran cantidad de circunstancias que condicionan esta naturaleza en el sonido producido por una persona, que van desde las fisiológicas hasta las culturales y del entorno. Es sorprendente que la onda acústica generada

por una persona al pronunciar una palabra, permita, por ejemplo, hacerse una idea aproximada de dónde vive o de dónde es originaria.

- **Las características cambiantes de la pronunciación.**

  El problema de la *variabilidad* en la pronunciación. La probabilidad de que una misma persona produzca dos ondas sonoras exactamente iguales al pronunciar una palabra en dos oportunidades, está muy próxima a cero. En esta categoría de problema se pueden mencionar infinidad de distintas causas: intensidad del sonido producido, conjunto de sonidos empleados, duración de los sonidos, entonación, hábitos de construcción de oraciones, ritmo de pronunciación, amplitud del vocabulario empleado, condición física y/o anímica de la persona en el momento de pronunciar.

1.2.3. Arquitectura de un sistema de reconocimiento de palabras aisladas

La mayoría de los sistemas automáticos de reconocimiento de palabras aisladas integran ciertos componentes o módulos fácilmente distinguibles dada su orientación específica a la solución de alguno de los subproblemas particulares presentes en el problema completo. No debe perderse de vista el hecho de que cada uno de estos componentes está asociado con todo un campo de investigación en particular, en los que equipos de desarrollo se encuentran participando activamente. La Figura 1.1 ilustra las distintas etapas que componen un sistema típico para el reconocimiento de palabras.

La primera etapa en el funcionamiento de un sistema de este tipo, consiste en la *adquisición* de una señal eléctrica que permita una representación adecuada de la onda sonora asociada con el fenómeno del que se realiza el estudio. En el caso de la voz humana, la presión del aire que transporta la onda acústica es sensada a través de un transductor que en muchos casos es un micrófono. El sonido de la voz que posee naturaleza analógica es entonces reemplazado a través del proceso de muestreo/conversión por valores discretos que pueden ser representados digitalmente con arreglos de ceros y unos en una computadora.

La segunda etapa presente en un sistema de reconocimiento de palabras consiste en la determinación de en qué parte de los datos obtenidos se encuentra la información que

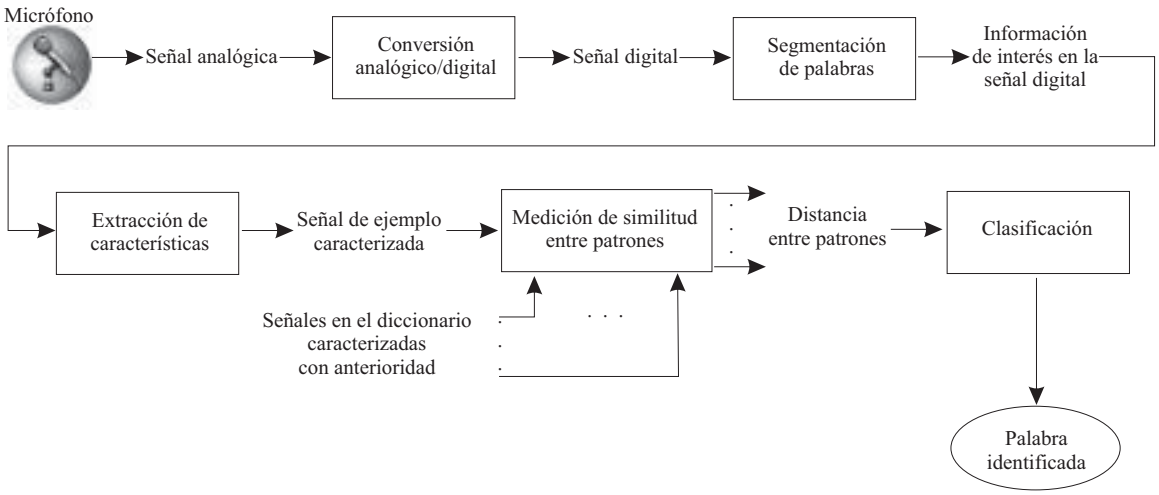

Figura 1.1: Estructura típica de un sistema de reconocimiento de voz.

se desea analizar. En otras palabras, debe determinarse dónde empieza y dónde termina la pronunciación de los elementos del lenguaje en los que se enfocará la atención; a esto se le denomina la *segmentación* de la información. El proceso de la producción de la voz humana incluye la emisión de sonidos simples denominados fonemas, o bien de sonidos compuestos de varios fonemas que se denominan bifonemas, trifonemas, etc. El problema que tiene una solución más simple en el renglón del aislamiento de la información que se estudia es la segmentación de palabras aisladas, donde se consideran todos los fonemas emitidos por un usuario y que, idealmente, están acotados por silencio previo y posterior.

La tercera etapa en el sistema de reconocimiento de palabras es la encargada de extraer características significativas de la señal adquirida, esto es *caracterizar* la información. El propósito de esta etapa es cambiar la representación original de la señal en una nueva representación más económica que la previa desde el punto de vista del volumen de datos requerido, con una forma tal que además destaque las diferencias entre dos señales diferentes y reafirme la similitud que se presuma entre señales parecidas o iguales. Esta etapa permite desechar la información redundante o de poco valor desde la perspectiva del reconocimiento del habla.

La cuarta etapa del funcionamiento del sistema que reconoce palabras se dedica a la comparación de dos señales caracterizadas en base a un mismo método, de manera que se

produzca alguna valoración que permita detectar cuando dos señales son similares o no lo son. Esto es lo que se denomina *comparación entre patrones* y por lo general se manifiesta como una medida de *distancia*. Es común en los desarrollos tecnológicos actuales, en los que el vocabulario de palabras a reconocer es grande, el contar con una colección de señales previamente caracterizadas y almacenadas (fuera-de-línea) e indexadas en un diccionario, contra las cuales se efectúa la comparación de una nueva señal que es adquirida al tener en funcionamiento el sistema (en-línea). De este modo, al efectuar el cálculo de las mediciones de las distancias entre los patrones de cada una de las señales conocidas a priori y patrón a prueba de una nueva señal adquirida, se proponen de entre los miembros en el diccionario un cierto número de candidatos viables de producir una identificación positiva al compararseles con el ejemplo de prueba.

La quinta etapa del sistema para reconocer palabras se dedica a la tarea de emitir un resultado final considerando las distancias arrojadas por la etapa de comparación entre patrones, a través de la adopción de algún criterio de decisión o de *clasificación*. En otras palabras, una vez que se ha determinado el patrón o los patrones más parecidos al comparar señales, se debe concluir que la nueva señal pertenece a alguna de las clases de señales conocidas por el sistema, o bien, podría determinarse que la nueva señal no pertenece a ninguna de las clases conocidas. De esta manera, si existe una decisión positiva, el sistema deberá asociar la señal procesada con los símbolos del lenguaje correspondientes, quizá escribiendo en pantalla o imprimiendo los caracteres que forman las palabras que el usuario pronunció.

1.2.4. Tipos de sistemas de reconocimiento de voz

Con el propósito de contar con parámetros objetivos, por ejemplo al efectuar la comparación entre sistemas automáticos de reconocimiento del habla, debe tenerse en consideración que la mayoría de tales sistemas se enfocan a la solución de alguno o algunos cuantos de los múltiples problemas que la tarea integral implica, esto debido a que no se cuenta con un único sistema que ofrezca una solución óptima para todos y cada uno de los inconvenientes que se pueden presentar.

A continuación se hace mención de algunas de las características de distintos de-

sarrollos elaborados por los especialistas en el área.

Respecto a la fluidez permitida en la pronunciación de las palabras que el sistema es capaz de reconocer:

- **Palabras aisladas:** el usuario debe pronunciar una sola palabra que se acota por los silencios que la preceden y suceden.
- **Discurso continuo:** el usuario puede pronunciar oraciones largas sin estar obligado a detener la emisión de sonidos entre palabra y palabra (no hay inserción de silencio).

Respecto al número de voces que el sistema es capaz de reconocer:

- **Dependiente:** el sistema es capaz de reconocer únicamente la voz de la persona para la cual ha sido entrenado.
- **Multiusuario:** el sistema reconoce un número limitado de voces de distintas personas quienes lo usen.
- **Independiente:** el sistema reconoce la voz de cualquier persona que lo use.

Es pertinente mencionar que un sistema de reconocimiento de voz no solo se caracteriza por operar sobre palabras aisladas o solo por ser de respuesta en tiempo real, por mencionar únicamente algunas de las diversas características que puede llegar a poseer, sino que su alta o baja calidad se integra con el equilibrio que se logre en conjunto entre todos los diferentes rubros a través de los cuales se le pueda llegar a evaluar. Es probable que se determine que un sistema es muy sobresaliente por su capacidad de reconocer discurso continuo, pero que se le catalogue como deficiente si es que esta capacidad hace que el tiempo de respuesta se extienda más allá de un cierto límite.

Respecto al volumen de palabras que el sistema es capaz de reconocer:

- **Vocabulario pequeño:** el sistema puede reconocer un número reducido de palabras, típicamente menos de 50.
- **Vocabulario mediano:** el sistema puede reconocer un número limitado de palabras, típicamente hasta 400.
- **Vocabulario grande:** el sistema puede reconocer un gran número de palabras, típicamente hasta 4,000. Con el manejo de un volumen de palabras como tal, se hace posible la construcción de sistemas de reconocimiento de dictado en un contexto de vocabulario restringido.
- **Vocabulario muy grande:** el sistema puede reconocer típicamente hasta 40,000 palabras.
- **Vocabulario ilimitado.**

Respecto al canal de transmisión de datos:

- **Local:** el micrófono con el que se capturan los datos está conectado a una computadora, que efectúa la conversión analógica/digital.
- **Remoto:** el micrófono con el que se capturan los datos está separado de la computadora, y los datos se transmiten analógicamente entre los puntos de emisión y recepción.

Respecto al tiempo de respuesta:

- **De tiempo real:** el sistema permite que se establezca un diálogo sin demoras, o bien puede iniciar una acción inmediatamente, en respuesta al reconocimiento de una orden o comando, por ejemplo “encender”, “apagar”, etc.
- **De largo tiempo de respuesta:** el establecimiento de un diálogo no es imperativo, o bien el reconocimiento de un probable comando puede permitir el transcurso de un tiempo adecuado antes de producir el inicio de una acción.

1.3. Objetivos de la Tesis

1.3.1. Objetivo general

En una perspectiva global, se desea aplicar la Transformada Wavelet Discreta (DWT, Discrete Wavelet Transform) como etapa de caracterización de señales en el problema de reconocimiento de palabras, módulo que se mencionó en la sección 1.2.3, y someter a escrutinio el desempeño de funciones Wavelet madre con distintas características, tratando de mantener invariante el entorno en el que se obtiene la señal que se intenta reconocer.

El problema completo del reconocimiento automático de la voz humana aún dista de estar resuelto de manera íntegra, puesto que en todos los componentes que forman un sistema de esta naturaleza se continúa con la búsqueda de alternativas a los métodos actualmente utilizados, de manera que se perfeccione el rendimiento que reportan.

Es por ello indispensable establecer acotaciones al problema específico que se pretende resolver con el presente desarrollo. El último de los objetivos particulares en esta sección señala el contexto en el que se puede ubicar la propuesta de este trabajo, de acuerdo a las características de los sistemas de reconocimiento de voz existentes que se mencionaron en la sección 1.2.4, sin embargo, la naturaleza del presente desarrollo no pretende impactar, mejorar o ampliar específicamente ninguna de tales capacidades. La pregunta esencial que se responde es si el reemplazo por una herramienta (para la cual se cuenta con evidencia de que se ajusta adecuadamente al análisis de seãles no estacionarias) alternativa a las existentes (Fourier, Cepstro, etc.) en uno solo de los m'odulos de un sistema de reconocimiento de voz, permite observar un desempeõ de tal sistema que aliente investigaciones futuras en esta misma línea.

1.3.2. Objetivos particulares

- Implantar la Transformada Wavelet Discreta mediante el algoritmo piramidal de Mallat, empleando como funciones base el Wavelet de Haar y los Wavelets de Daubechies de ordenes 2 a 10.

- Integrar la Transformada Wavelet Discreta en un sistema autónomo, junto a los módu-

los necesarios para abordar los restantes aspectos del problema completo de reconocimiento automático de palabras.

- Evaluar el rendimiento del sistema haciendo uso de la *Teoría de Detección de Señales* elaborando las curvas ROC (*Receiver Operating Characteristic*) correspondientes, como medio para la elección de un modelo posiblemente óptimo y descartar otros posiblemente subóptimos.

- El sistema a construir permitirá reconocer palabras aisladas, será dependiente del usuario, con tiempo de respuesta aceptable, relativa buena certeza de reconocimiento para el usuario quien lo entrenó al ser usado en entornos de baja interferencia acústica.

1.4. Descripción de Capítulos

1.4.1. Organización de este documento

El resto de este documento guarda la organización que enseguida se muestra.

El capítulo 2 comprende la descripción de algunas técnicas y algoritmos empleados de manera cotidiana en el ámbito de la caracterización de señales. Se hace mención del tratamiento aplicado a la solución del problema de la segmentación de palabras en la señal. Se hace especial énfasis en la utilización de la Transformación con Wavelets, particularmente útil en el análisis de señales que raramente son estacionarias, como lo es el caso de la voz humana. Se presenta asimismo, la implantación de la Transformada Wavelet Discreta mediante algoritmos piramidales.

El capítulo 3 contiene la descripción de algunos algoritmos utilizados comunmente para llevar a cabo la tarea de medir la distancia o similitud entre patrones. Del mismo modo, se describe la implantación que se llevó a cabo del esquema de Doblado Dinámico en Tiempo (DTW, Dynamic Time Warping). Asimismo se describen los criterios de clasificación implementados para la emisión de un resultado final por parte del sistema.

El capítulo 4 incluye los detalles de implantación en código de los algoritmos necesarios para la construcción de un sistema básico pero completo, para el reconocimiento de palabras pronunciadas aisladamente.

El capítulo 5 plasma los resultados obtenidos al contrastar el desempeño de la etapa de caracterización de la señal, al emplear dos familias de Wavelets diferentes en la transformación.

Finalmente en el capítulo 6 se revelan las conclusiones que se derivaron del presente desarrollo, y se establecen indicadores de desarrollo futuro.

Capítulo 2

El procesamiento digital de la señal de voz

Lo que somos capaces de percibir respecto de la producción de la voz humana es una cantidad física de naturaleza analógica. El análisis numérico sin embargo, se favorece con una representación digital a la que denominamos señal. Este capítulo está dedicado a mostrar el tratamiento de la señal de la voz humana en forma digital, con objeto de reconocer de ella las características distintivas que posee.

Se presentan los esquemas más populares que se emplean para la determinación del inicio y el fin de una palabra, las técnicas tradicionales de caracterización de señales acústicas, y se lleva a cabo la implantación de la Transformada Discreta Wavelet basada tanto en el Wavelet de Haar como en el Wavelet de Daubechies, este último de diversos ordenes.

2.1. Segmentación de la señal de voz

Uno de los problemas que inicialmente aparecen al procesar la señal de la voz es que, debido a la naturaleza continua del discurso de las personas, la señal acústica generalmente no se interrumpe muy a menudo. Cuando una persona lee un texto, el espacio ortográfico que separa las palabras no siempre se traduce en "silencios" claramente definidos

que separen los distintos elementos del lenguaje que se emiten como sonidos.

Un enfoque que permite evitar tal problema, consiste en obligar al usuario del sistema a guardar silencio cada vez que ha terminado de pronunciar una palabra. Podría pensarse que una señal obtenida bajo esta condición favorece la consecución de buenos resultados al analizarla, ya que se podría tener la certeza de que "en algún lugar" entre el inicio y el fin de los datos digitalizados se encuentra la información que representa la palabra pronunciada.

A continuación se describen las características que se extraen de la señal para el efecto de llevar a cabo la segmentación.

2.1.1. Energía de tiempo corto

La *energía de tiempo corto* E para el segmento $[a, b]$ de la señal se determina por la ecuación 2.1.

$$E = \sum_{i=a}^{b} \left[x\left(i\right)^2 \right] \tag{2.1}$$

donde $x(i)$ es la i-ésima muestra en la señal.

Si el reconocimiento de la voz se efectúa en un ambiente de baja contaminación acústica, entonces el mayor contenido energético en la señal representada por los datos a procesar se tiene cuando la persona está pronunciando una palabra. De esta manera, antes de que la persona inicie la pronunciación, así como también después de hacerlo, el contenido energético será relativamente menor. Si se efectúa el cálculo de la energía en segmentos breves de la señal y luego se comparan los resultados, se obtiene una estimación de la amplitud relativa de la señal entre los segmentos comparados. De esta manera, cuando la *energía de tiempo corto* en un segmento de la señal rebasa cierto umbral con respecto a los segmentos previos, puede inferirse que la persona ha comenzado a pronunciar. Un razonamiento similar se usa para la detección del fin de una palabra.

2.1.2. Magnitud promedio de tiempo corto

La necesidad de calcular el cuadrado de cada muestra en el segmento en cuestión, puede resultar en una estrategia demasiado costosa desde el punto de vista computacional. Una alternativa útil es la que se menciona enseguida.

La *magnitud promedio de tiempo corto* M para el segmento $[a, b]$ de la señal se determina por la ecuación 2.2.

$$M = \sum_{i=a}^{b} |x\,(i)| \tag{2.2}$$

donde $x(i)$ es la i-ésima muestra en la señal.

La comparación de la amplitud relativa entre dos segmentos de una señal puede obtenerse también si en lugar de efectuar la suma de los cuadrados de las magnitudes de las muestras en un segmento, se realiza la suma de los valores absolutos de las muestras en el segmento, eliminándose así la necesidad de elevar un número a una potencia.

2.1.3. Régimen de cruces por cero de tiempo corto

El *régimen de cruces por cero de tiempo corto* Z para el segmento $[a, b]$ de la señal se determina con la ecuación 2.3.

$$Z = \sum_{i=a}^{b} |\text{signo}\,[x\,(i)] - \text{signo}\,[x\,(i-1)]| \tag{2.3}$$

donde $x(i)$ es la i-ésima muestra en la señal.

Se puede asociar el valor del régimen de cruces por cero de tiempo corto con una estimación del valor promedio de la frecuencia de los datos en la señal para el segmento en cuestión.

Donde:

$$\text{signo}\,[x\,(i)] = \begin{cases} 1 \;\forall\; x\,(i) > 0 \\ 0 \;\forall\; x\,(i) = 0 \\ -1 \;\forall\; x\,(i) < 0 \end{cases} \tag{2.4}$$

Antes que el sistema de reconocimiento de voz comience a recibir los datos correspondientes a la palabra pronunciada por el usuario, las perturbaciones llevarán un bajo contenido en frecuencia, siempre que se mantengan los límites de baja contaminación acústica

requeridos. Esto ocurre de la misma forma cuando el usuario ha terminado de pronunciar. Sin embargo, a lo largo de la pronunciación el contenido de frecuencias en la señal fluctúa considerablemente, de manera que la naturaleza oscilatoria de las ondas sonoras provoca que la señal pase en rápida sucesión de adquirir valores positivos a poseer valores negativos y viceversa a medida que el tiempo transcurre. Esto implica que necesariamente, la señal cruza por cero y la presencia de gran cantidad de estos cruces por unidad de tiempo, así como un comportamiento consistente en este sentido a lo largo de varios segmentos de la señal, constituyen un fuerte indicio de que se ha iniciado la pronunciación de una palabra. Un razonamiento similar se usa para la detección del fin de una palabra.

2.2. Transformadas de Fourier y de Fourier de tiempo corto

2.2.1. Transformada de Fourier

Una de las herramientas en el vasto arsenal disponible para el procesamiento digital de señales es la Transformada de Fourier Continua (CFT, Continuous Fourier Transform) de la función x donde esta última depende de la variable continua t. Esta es la herramienta ideal para estudiar una señal estacionaria, cuyas propiedades promedio no varían con el tiempo. En otras palabras, las señales estacionarias se descomponen canónicamente en combinaciones lineales de ondas (senos y cosenos) [Meyer93].

La amplitud y la fase de una componente sinusoidal de $x(t)$ depende de la frecuencia de tal componente; en términos de la frecuencia ordinaria (f), la CFT se encuentra definida por la ecuación 2.5.

$$X(f) = \int_{-\infty}^{+\infty} x(t) \cdot e^{-i2\pi ft} dt \tag{2.5}$$

Para la utilización de esta herramienta en computadoras se puede obtener una función de "tiempo discreto" muestreando una función de tiempo continuo $x(t)$, la cual produce una secuencia $x(nT)$ para valores enteros de n. La Transformada de Fourier Discreta (DFT, Discrete Fourier Transform) es equivalente a la CFT de una función "continua" que se construye usando la secuencia $x[n] = x(nT)$ para modular un tren de impulsos. La DFT

se determina mediante la ecuación 2.6.

$$X_T(f) = \sum_{n=-\infty}^{+\infty} x[n] \cdot e^{-i2\pi f n T} = \sum_{n=-\infty}^{+\infty} x[n] \cdot e^{-i2\pi \frac{f}{f_s} n} \tag{2.6}$$

la cual es una función periódica con período $f_s = 1/T$. Un punto de vista alternativo es el de que la DFT es una transformación al dominio de la frecuencia que está acotada (o es finita), con extensión f_s.

La utilidad de la DFT en el ámbito del procesamiento de señales, radica en el hecho de que se le puede ofrecer una señal en el dominio del tiempo para que efectúe un análisis en búsqueda de su contenido en frecuencia. El resultado de la aplicación de esta técnica sobre una señal del dominio del tiempo es una serie de coeficientes que representan la contribución de las funciones base senoidales en cada frecuencia. La Figura 2.1 muestra como ejemplo a la izquierda las representaciones de tres funciones sinusoidales puras en el dominio del tiempo registradas durante un intervalo de 1 s, mientras que a la derecha de cada una de ellas se observa su representación equivalente en el dominio de la frecuencia, toda vez que se les ha sometido a la acción de la DFT. Los picos representan la magnitud de los números complejos que entrega la transformación. La primera de estas funciones que es $cos(2\pi * 10t)$ posee una frecuencia de 10 hz, la segunda función que es $cos(2\pi * 50t)$ tiene una frecuencia de 50 hz, y la última función es $cos(2\pi * 100t)$ con una frecuencia de 100 hz. La Transformada de Fourier revela con toda exactitud el contenido de frecuencia en las señales.

2.2.2. Desventajas en el uso de la DFT para el análisis de la señal de voz

Al tratar una señal que presente un comportamiento estacionario, si se cuenta con suficiente información a priori es factible predecir de manera confiable cual será el comportamiento de la señal en un tiempo futuro. Esto puede interpretarse también en el sentido de que los contenidos de frecuencia en una señal no varían, o si lo hacen, ello ocurre de una manera sistemática.

La Figura 2.2 muestra a la izquierda tres señales distintas entre sí. La primera de

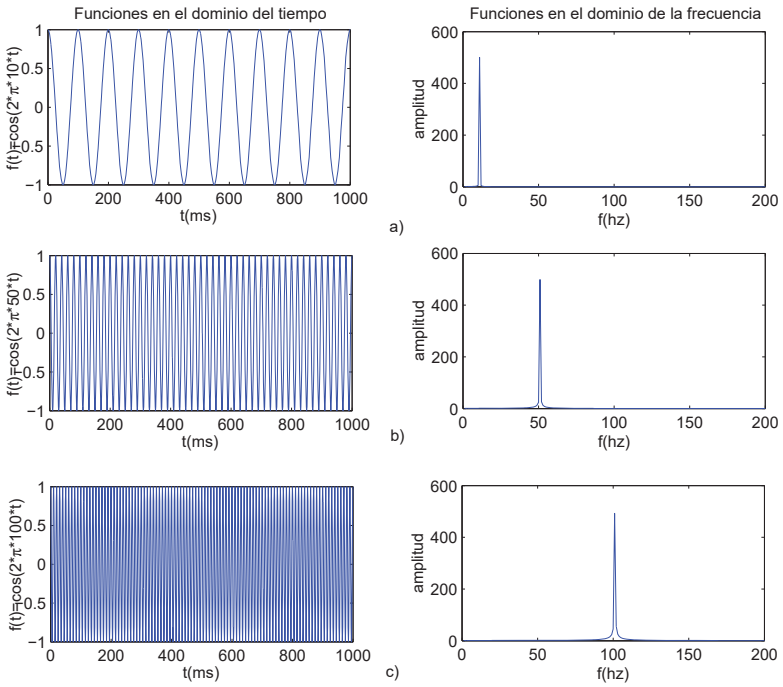

Figura 2.1: Señales sinusoidales puras (a la izquierda). Espectros de magnitudes de las DFT correspondientes a las señales sinusoidales puras (a la derecha).

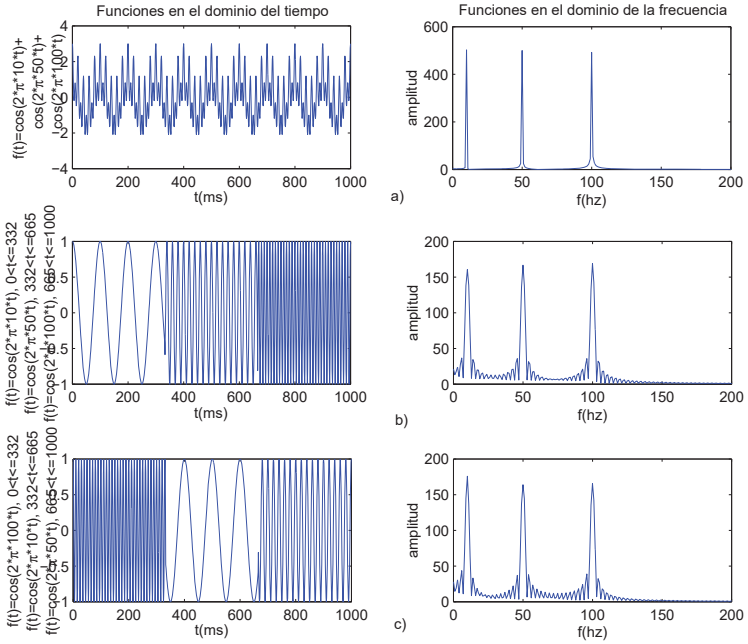

Figura 2.2: Una señal estacionaria y dos señales no estacionarias (a la izquierda). Espectros de magnitudes de las DFT correspondientes a la señal estacionaria y a las dos señales no estacionarias (a la derecha).

ellas es estacionaria,

$$f(t) = cos(2\pi * 10t) + cos(2\pi * 50t) + cos(2\pi * 100t) \tag{2.7}$$

mientras las dos restantes no lo son:

$$\begin{aligned} f\left(t\right) &= cos\left(2\pi * 10t\right), 0 < t \leq 332 \\ f\left(t\right) &= cos\left(2\pi * 50t\right), 332 < t \leq 665 \\ f\left(t\right) &= cos\left(2\pi * 100t\right), 665 < t \leq 1000 \end{aligned} \tag{2.8}$$

$$\begin{aligned} f\left(t\right) &= cos\left(2\pi * 100t\right), 0 < t \leq 332 \\ f\left(t\right) &= cos\left(2\pi * 10t\right), 332 < t \leq 665 \\ f\left(t\right) &= cos\left(2\pi * 50t\right), 665 < t \leq 1000 \end{aligned} \tag{2.9}$$

A la derecha en la Figura 2.2 se observan los correspondientes espectros de las magnitudes de las DFTs de cada una de las señales en la izquierda. Si se desprecian las diferencias en amplitud y los rizos de magnitudes relativamente pequeñas que aparecen en la segunda y tercera transformadas (a los que se conoce con el nombre de *"spectral leakage"* o *"escurrimiento espectral"*) [Smith97], el aspecto de las tres señales en el dominio de la frecuencia es el mismo. El procesamiento mediante la Transformada de Fourier de tales señales con diferencias tan radicales en el dominio del tiempo, produce una única representación en el dominio de la frecuencia, hecho que hace imposible efectuar distinción alguna entre las señales originales en el dominio del tiempo a partir de las cuales se genera esa última representación. El escurrimiento entre los picos de magnitudes altas se presenta por las transiciones abruptas de frecuencias, al perturbarse la "suavidad" de cambio de las funciones sinusoidales.

Infortunadamente, la mayor parte de las señales producidas por fenómenos físicos que podría interesarnos estudiar, como lo son las señales generadas por eventos acústicos, no presentan un comportamiento estacionario, siendo pues, señales que no cumplen con la definición de las estacionarias y para las que no se puede predecir cómo será su comportamiento en el futuro a pesar de contar con un muy alto conocimiento acerca de su evolución pasada. En la práctica, entonces, los fenómenos acústicos o de vibración raramente son absolutamente estacionarios. En algunos casos la variación ocurre de manera suficientemente

lenta dentro de un período breve de muestreo, de manera que se le pueda considerar como cuasi-estacionaria o aproximarse a una señal estacionaria en porciones pequeñas o en tiempo corto [Hansen97].

Finalmente, y representando quizá la carencia más significativa que demuestra la DFT, se conoce el hecho de que pierde por completo la información *temporal* asociada con los eventos individuales en la señal a la cual se aplica, de manera tal que se hace imposible determinar *cuándo* estos ocurren y en qué *orden* se presentan. En otras palabras, la DFT no posee capacidad alguna de resolución en tiempo a cambio de ofrecer una resolución perfecta en frecuencia, esto último siempre y cuando la señal sea estacionaria, lo que significa que se puede conocer con absoluta precisión el contenido de frecuencia en una señal, pero no es posible ubicar en el tiempo el momento en que tales componentes aparecen.

La Figura 2.2 ilustra también con toda claridad la dificultad causada por la pérdida de la información temporal. Si se observan las representaciones en el dominio del tiempo para la segunda y tercera funciones 2.8 y 2.9 respectivamente, debe notarse que al inicio de su manifestación 2.8 posee un contenido de baja frecuencia, mientras que para 2.9 en el mismo período de tiempo se aprecia un contenido de alta frecuencia. Al revisar la representación de estas dos funciones en el dominio de la frecuencia, se encuentra que para el inicio de la manifestación de 2.8 corresponde el primer pico de frecuencia que se observa de izquierda a derecha en la gráfica (el que denota la frecuencia más baja de 10 hz); para 2.9 el pico de frecuencia correspondiente al primer período de su manifestación es el tercero (el que se aprecia en la extrema derecha de la gráfica, el asociado a la frecuencia más alta de 100 hz). Solo el contar con ambas representaciones, la del dominio del tiempo y la del dominio de la frecuencia simultáneamente nos permite efectuar estas afirmaciones.

La DFT no conoce ningún detalle de la representación en el dominio del tiempo, y es por ello incapaz de informar de la ubicación, el orden y la duración de eventos.

2.2.3. La Transformada de Fourier Discreta de Tiempo Corto

La Transformada de Fourier Discreta con Ventanas, también conocida como Transformada de Fourier Discreta de Tiempo/Término Corto [Rabiner78] (STFT, Short Time/Term Fourier Transform) presenta una alternativa que ofrece mejores resultados cuando

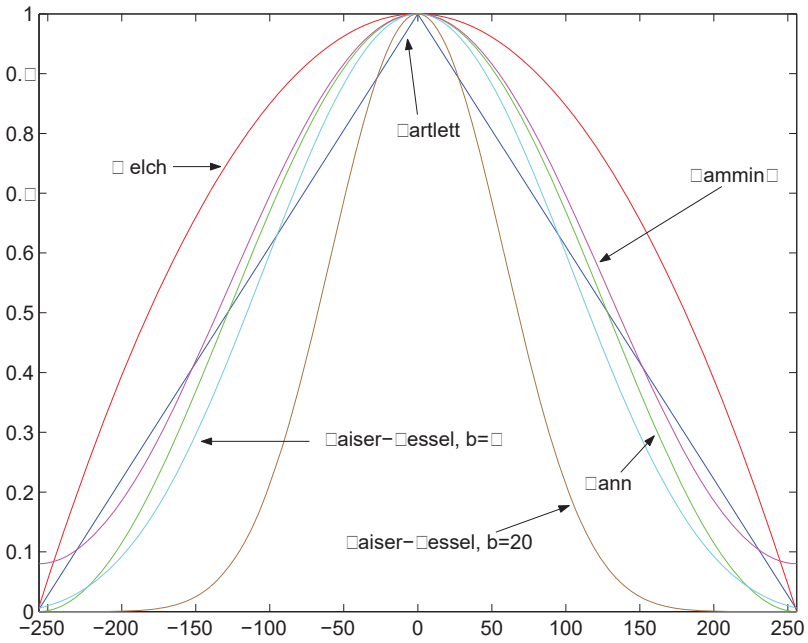

Figura 2.3: Tipos de ventanas empleadas para seccionar señales.

se realiza el análisis de señales que no son estacionarias. Cuando se utiliza esta herramienta, la señal a ser transformada es "cortada" en secciones y entonces se aplica la DFT a cada una de estas secciones en la búsqueda de los contenidos de frecuencia por separado.

En virtud de que muy probablemente las transiciones en los límites de las secciones practicadas a la señal serán agudas, tal como ocurrió con los cambios entre frecuencias en las funciones 2.8 y 2.9, se emplea una ventana localizada en tiempo o con soporte compacto que secciona los datos de entrada y forza a que los extremos de la sección converjan a cero. La función ventana que se elige es por lo tanto una que enfatiza más los datos al centro de la ventana que los datos en los extremos de la misma. Además es muy frecuente el que se permita un ligero traslape entre dos ventanas consecutivas, de manera que las transiciones son aún menos abruptas. El soporte y la forma de la ventana seleccionada se mantiene invariante durante todo el análisis, es decir, no se le puede cambiar por otra en el transcurso del estudio. Los rizos en la Figura 2.2 representan frecuencias inexistentes en la señal del dominio del tiempo. Este tipo de perturbaciones son las que pueden aparecer al seccionar la señal, por ejemplo, con una ventana de forma rectangular, así que la forma de

la ventana debe ser tal que atenúe las frecuencias con alto valor de escurrimiento, las que aparecen próximas a las frecuencias realmente existentes en la señal y poseen magnitudes relativamente altas.

Algunas de las ventanas que se emplean para cortar las señales en este tipo de transformación son: a) Cuadrada, b) de Welch, c) de Bartlett d) de Hann, e) de Hamming, f) de Kaiser-Bessel [Smith97]. El aspecto característico de las ventanas se observa en la Figura 2.3.

El efecto de seccionar la duración de la señal mediante una ventana es el de localizar a la señal en el tiempo, de manera que ahora se cuenta con información acerca de su contenido de frecuencia, así como también se sabe si un contenido de frecuencia aparece al inicio, a la mitad o al final de la manifestación de la señal. Sin embargo, la resolución que se puede alcanzar con la aplicación de esta técnica está determinada por la localización en tiempo que tenga la ventana elegida. Para modificar esta resolución se hace necesario un nuevo análisis completo utilizando una ventana con distinta localización en el tiempo que la empleada en el análisis previo.

Un caso extremo se presenta al utilizar una ventana de duración infinita y, por lo tanto, sin localización en el tiempo. En estas circunstancias, se obtiene el mismo funcionamiento de la Transformada de Fourier sin el uso de ventanas.

El otro caso extremo se tiene cuando la ventana que se utiliza está sumamente localizada en el tiempo, circunstancia que arroja como resultado una mejor resolución en el tiempo sacrificándose la resolución en frecuencia. Ahora se sabe con precisión donde aparecen los componentes de frecuencia, pero no se puede precisar con mucha exactitud cuál es el aporte de tales contenidos de frecuencia en la nueva representación de la señal. Esto sugiere que el producto entre la magnitud de un intervalo de frecuencia y la magnitud de un intervalo de tiempo es una cantidad estable.

Estos dos casos opuestos ilustran con claridad el cumplimiento del principio de incertidumbre (o de indeterminación) de Heissenberg aplicado a las propiedades de las señales [Strang97].

Resulta entendible pues, que deberá elegirse una ventana con una localización en el tiempo que constituya un buen compromiso entre la resolución deseada en tiempo y la

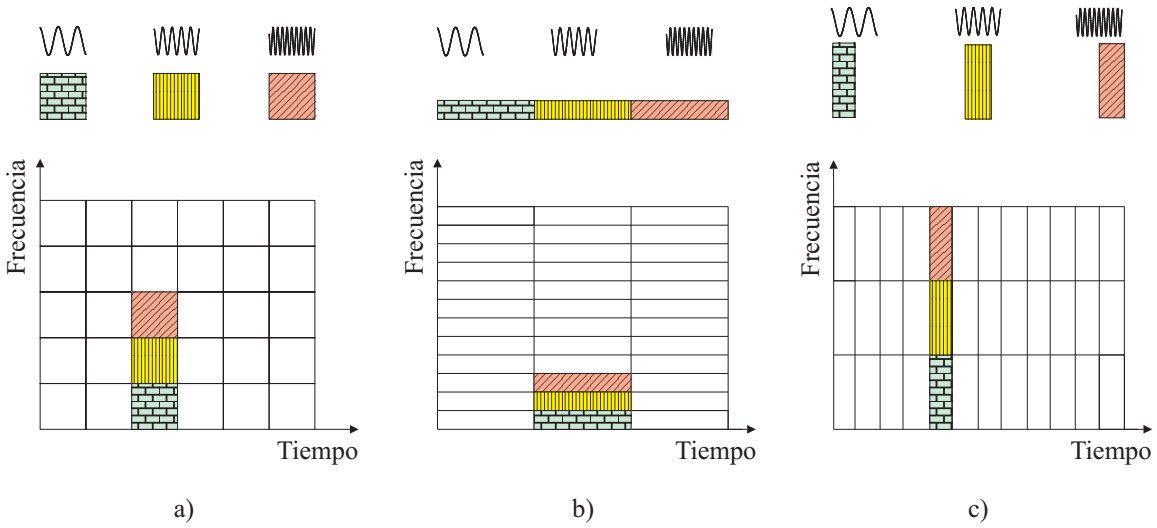

Figura 2.4: El análisis de Fourier de Tiempo Corto ofrece una resolución fija tanto en tiempo como en frecuencia.

resolución deseada en frecuencia. Ambas resoluciones pueden interpretarse como una única resolución combinada que se obtendrá al estudiar una señal, debido a que la misma ventana debe ser utilizada para el análisis completo.

Además del análisis de Fourier con resolución variable existe una gran cantidad de metodologías de análisis espectral para señales, entre las que se cuentan: Codificación por transformación, Predicción lineal selectiva, Vocoder de banda base, Codificación por bandas de frecuencia, Modelado con resolución variable y Modelado autoregresivo generalizado.

En la Figura 2.4 se muestra una ilustración para ayudar a comprender el efecto que tiene la elección del soporte o localización en tiempo de la ventana utilizada en el análisis de Fourier de Tiempo Corto. En la parte superior de la imagen se observan las funciones base del análisis cuya frecuencia se incrementa de izquierda a derecha y a las que se asocia un recuadro como los que contiene el gráfico en el plano cartesiano. Una vez en el plano, la función base de menor frecuencia aparece más abajo en el eje de las ordenadas, mientras que la función base de frecuencia más alta aparece arriba. Los pequeños mosaicos que representan cada uno a una función base, tienen la misma longitud en su dimensión paralela al eje de la frecuencia, y esto representa la equivalencia del soporte de la ventana en su capacidad de producir un resultado que ayude a discernir contenidos en frecuencia.

De la misma forma, los mosaicos tienen la misma longitud en su dimensión paralela al eje del tiempo, y esto denota la equivalencia del soporte de la ventana en su capacidad de informar sobre localización en el tiempo. Esta resolución se mantiene constante para el estudio completo [Proakis92].

2.3. Wavelets

La mayoría de las señales interesantes de ser analizadas contienen una gran cantidad de efectos transitorios o no estacionarios: caídas, escaladas, cambios abruptos, e inicios y finales de eventos. Frecuentemente estas características constituyen la parte más importante en la señal. Estas señales requieren un enfoque de análisis más versátil que el Análisis de Fourier, uno en el cual se tenga la posibilidad de hacer variar el tamaño de las ventanas que seccionan la señal, con el objeto de poder determinar con mejor precisión bien la información temporal o la información de contenido en frecuencia, aquello que sea lo más relevante. Los Wavelets son funciones matemáticas que satisfacen los requisitos de ser oscilatorias, localizadas en el tiempo, con valor efectivo igual a cero y son utilizadas en la representación de otras funciones o señales [Meyer93, Strang97].

2.3.1. Análisis con Wavelets

La técnica de Wavelets permite un análisis con el enfoque requerido en el estudio de señales no estacionarias, en el que se utiliza la técnica de seccionamiento de la señal mediante ventanas de tamaño variable.

El análisis con Wavelets admite el empleo de ventanas que abarcan intervalos de tiempo largos, si lo que nos ocupa es la determinación precisa de los contenidos en frecuencia baja, mientras que se pueden emplear ventanas que abarcan espacios de tiempo cortos cuando la preocupación específica sea determinar los contenidos de frecuencia alta [Graps95, Daubechies96].

La Figura 2.5 muestra la versatilidad en resolución que ofrece el uso de las funciones base en el análisis con Wavelets. En la parte superior de la imagen se observan las funciones base del análisis, cuya frecuencia se incrementa de izquierda a derecha y a las que se asocia

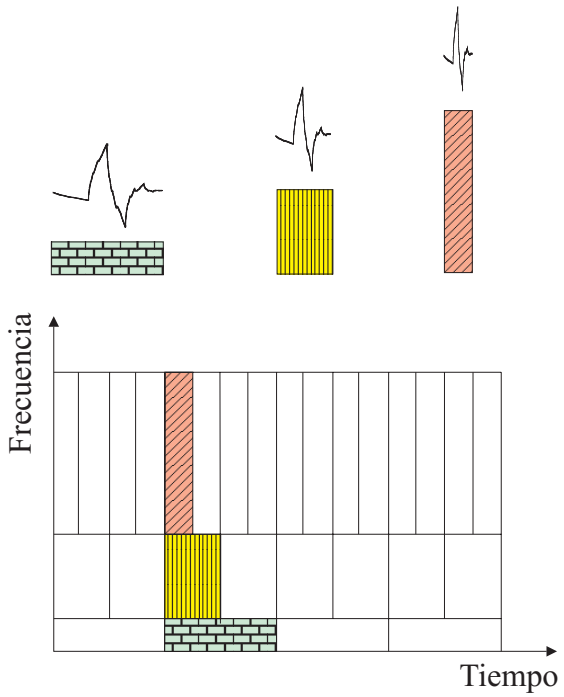

Figura 2.5: El análisis con Wavelets ofrece una resolución variable tanto en frecuencia como en tiempo.

un rectángulo como los que contiene el gráfico en el plano cartesiano. Una vez en el plano, la función base de menor frecuencia aparece más abajo en el eje de las ordenadas, mientras que la función base de frecuencia más alta aparece arriba. Los pequeños mosaicos que representan cada uno a una función base, tienen longitud creciente en su dimensión paralela al eje de la frecuencia cuando esta se incrementa, y esto representa que mientras más baja es la frecuencia de la función base, es mejor su capacidad de discernir contenidos en frecuencia. Por otra parte, los mosaicos tienen longitud decreciente en su dimensión paralela al eje del tiempo cuando su frecuencia se incrementa, y esto denota que mientras más alta es la frecuencia de la función base mejor es su capacidad de informar sobre localización en el tiempo. La resolución, por tanto, se mantiene variable a lo largo del estudio completo, lo cual puede entenderse como un único análisis multiresolución o con resolución múltiple.

Un Wavelet es una forma ondulada que posee una duración limitada y tiene un valor promedio igual a cero. A diferencia de los Wavelets, las sinusoides que son las funciones base en el análisis de Fourier no tienen duración limitada, es decir, se extienden desde $-\infty$

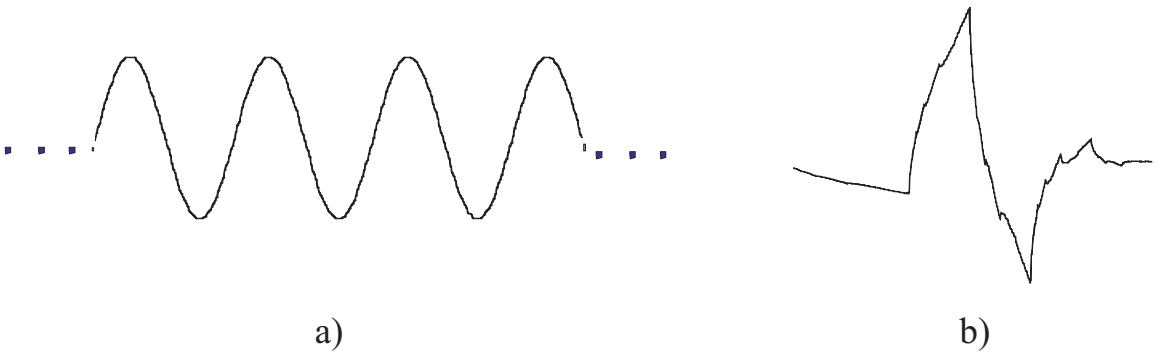

Figura 2.6: a) Funciones base del análisis de Fourier. b) Funciones base del análisis con Wavelets (en la gráfica, el Wavelet de Daubechies de orden 2).

hasta $+\infty$ a lo largo del tiempo. Mientras que las sinusoides tienen formas de onda suavizadas y son altamente predecibles, los Wavelets pueden parecer irregulares y asimétricos. La Figura 2.6a) muestra la forma característica de las formas de onda de las funciones base del análisis de Fourier, que poseen duración infinita, mientras que la Figura 2.6b) describe una de las formas de onda para una posible función base del análisis con Wavelets, cuyo soporte es compacto, lo cual significa que en sus extremos la función decae inexorablemente hasta anularse.

De una manera similar al análisis de Fourier, en el que una señal se descompone en diversas ondas sinusoidales de distintas frecuencias y amplitudes, en el análisis con Wavelets se descompone la señal en versiones escaladas y trasladadas del Wavelet originalmente elegido, al que se denomina el Wavelet de análisis o Wavelet madre.

2.3.2. Transformada Wavelet Continua

La Transformada Wavelet Continua (CWT, Continuous Wavelet Transform) [Cohen96] se define como la suma para todo tiempo t de la señal bajo estudio f multiplicada por versiones escaladas y trasladadas de la función Wavelet Ψ, hecho establecido por la ecuación 2.10.

$$C\left(escala,\ posición\right) = \int_{-\infty}^{+\infty} f\left(t\right) \Psi\left(escala,\ posición,\ t\right)\ dt \tag{2.10}$$

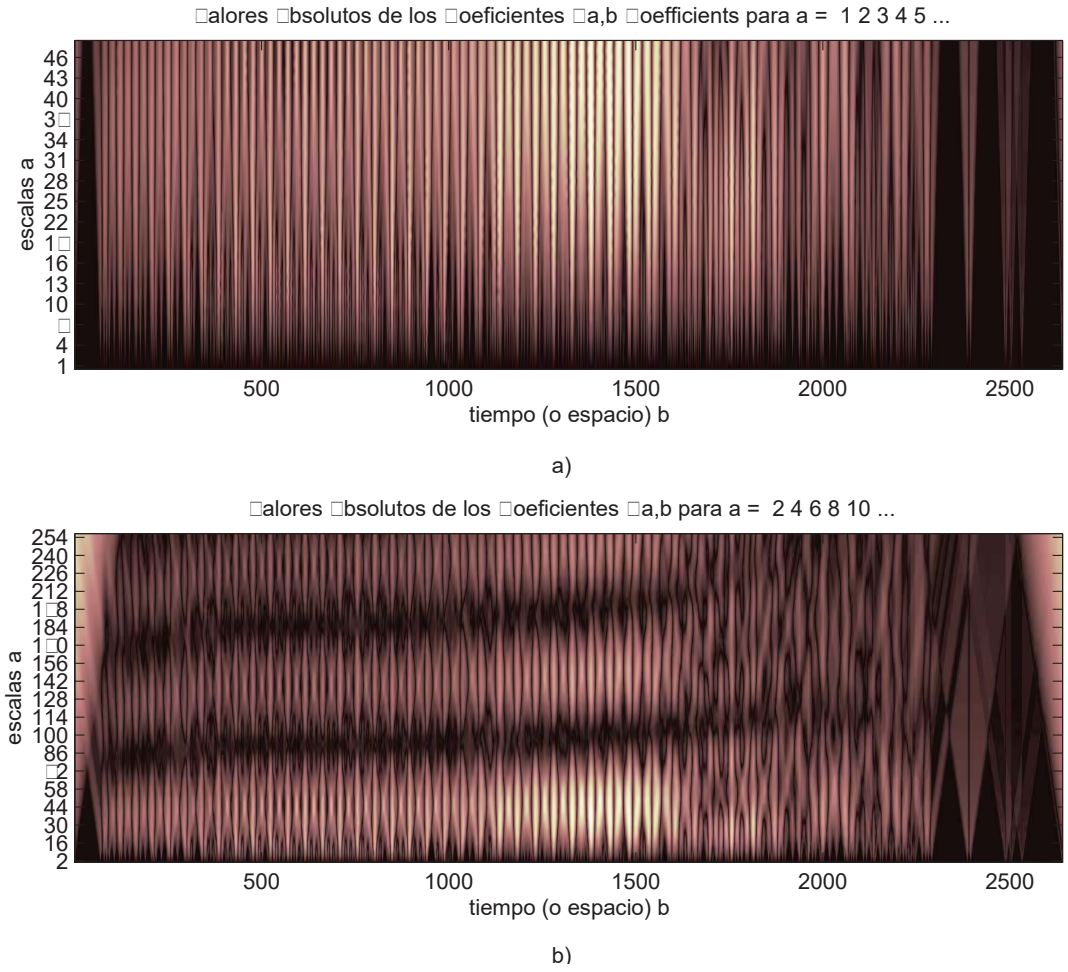

Figura 2.7: Coeficientes Wavelet para el análisis en escala continua para una elocución de la palabra *"uno"* sobre cada escala entre 1 y 48 (arriba), y para el análisis en escala continua de la misma palabra sobre cada segunda escala entre 2 y 256 (abajo).

El resultado de la aplicación de esta operación sobre la señal, son muchos coeficientes Wavelet C, que son función de las variables *escala* y *posición*. En otras palabras, la nueva representación de una señal en términos de coeficientes Wavelet es una vista de los datos en la que se cuenta simultáneamente con información acerca de la escala, que guarda cierta relación con el contenido en frecuencia, y también se cuenta con información acerca de la posición, que en el caso de una señal unidimensional como lo es la onda acústica, está relacionada con el tiempo en que se presentan los eventos cuando se pronuncia una palabra.

Aún cuando el análisis numérico automatizado en dispositivos digitales requiera necesariamente que los datos presenten una vista discretizada de los fenómenos físicos, en este caso de la señal acústica y que, por lo tanto, las transformaciones definidas para funciones continuas deban ajustarse a esta necesidad al ser implantadas en algoritmos, en el caso de la Transformada Wavelet se presenta la circunstancia particular de que al ser los coeficientes una función de las variables *escala* y *posición*, se puede efectuar la variación de la primera de ellas (la escala) sobre cualquier valor entero dentro del rango para ella establecida. Esto es lo que se conoce como la Transformada Wavelet Continua, que se distingue de la Transformada Wavelet Discreta en el hecho de que esta última no permite la elección de cualquier valor de escala dentro de su rango, sino solamente ciertas escalas particulares que se mencionan en la sección 2.3.3.

La Figura 2.7 muestra una posible interpretación del resultado de aplicar la Transformada Wavelet Continua sobre una señal. En las dos gráficas se muestra en cada punto el valor absoluto de un coeficiente Wavelet al estudiar una elocución de la palabra *"uno"*. Las zonas de más alta luminosidad indican la presencia de valores absolutos altos en los coeficientes resultantes. La primera gráfica es el resultado del cálculo de coeficientes incrementando la escala en 1 en cada ocasión, desde 1 hasta 48. La segunda gráfica es el resultado del cálculo de coeficientes incrementando la escala en 2 en cada ocasión desde 2 hasta 256.

La multiplicación de cada coeficiente obtenido con el Wavelet escalado y trasladado apropiadamente arroja como resultado los Wavelets que constituyen la señal original.

2.3.2.1. Escala y análisis multiresolución

En el análisis con Wavelets la *escala* que empleamos para observar los datos bajo estudio cobra una relevancia notable. Esta noción de escala, la cual claramente se refiere a la cartografía, implica que la señal se reemplaza, en una escala dada, por la mejor posible aproximación que puede esbozarse en esa escala. "Deslizándose" desde las escalas grandes hacia las escalas finas se efectúa un "acercamiento" y se llega cada vez a representaciones más y más exactas de la señal. Los algoritmos que se basan en Wavelets procesan datos en diferentes escalas o *resoluciones*. Cuando se aprecian las señales a través de una ventana de larga duración se distinguirán las características "gruesas" o persistentes, mientras que al

apreciar la misma señal a través de una ventana de corta duración serán más evidentes las características "finas" o de duración breve.

La parte del análisis que permite la identificación de las características gruesas lo cual se asocia con una escala alta, se realiza utilizando una versión dilatada o de baja frecuencia del Wavelet madre. En contraste, la parte del análisis que permite la identificación de las características finas lo cual se asocia con una escala baja, se realiza utilizando una versión contraída o de alta frecuencia del Wavelet madre. Por lo tanto, en el contexto del análisis multiresolución "escalar" un Wavelet significa bien "estirarlo" o por el contrario, "comprimirlo".

Una descripción más formal del concepto de escala se logra mediante la introducción de un factor de escala a, que indique con claridad cómo es que distintas versiones de un Wavelet se comparan. Mientras más pequeño es el factor de escala tanto más "comprimido" es el Wavelet.

La Figura 2.8 muestra tres versiones del mismo Wavelet madre: de Daubechies de orden 2 en tres escalas distintas. La función que describe la curva en 2.8a) tiene la forma: $f(t) = \Psi(t)$, donde el factor de escala se puede denotar con $a = 1$. La descripción matemática para la curva en 2.8b) es: $f(t) = \Psi(2t)$, de donde se desprende que el factor de escala es ahora $a = \frac{1}{2}$. Finalmente, la tercera versión del Wavelet en 2.8c) puede describirse con $f(t) = \Psi(4t)$, donde el factor de escala es $a = \frac{1}{4}$.

Es claro ahora que en el análisis multiresolución, la escala está relacionada inversamente con la frecuencia de la señal. Si las escalas altas corresponden con las versiones más "estiradas" del Wavelet madre, entonces mayor será la porción de la señal abarcada por la extensión del Wavelet y más gruesas son las características destacadas en la señal. Así, la correspondencia entre escala de Wavelets y contenido de frecuencia en la señal se establece en la Tabla 2.1.

2.3.2.2. Traslación

Los coeficientes Wavelet que se obtienen de la transformación son función de la escala y la posición. La traslación o el "deslizar" un Wavelet significa simplemente retrasar o adelantar su aparición. Matemáticamente, el retraso de una función se representa por

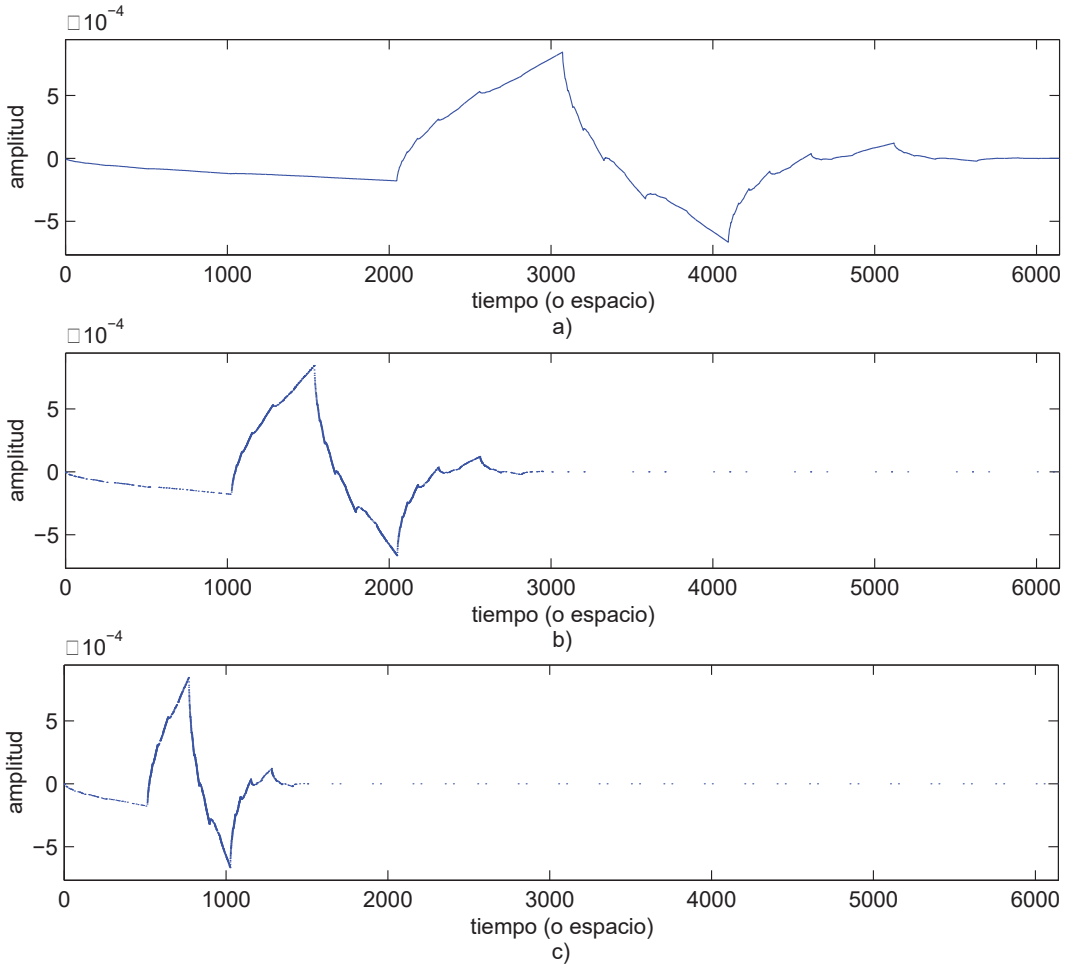

Figura 2.8: Wavelet madre de Daubechies de orden 2 en tres escalas distintas.

Tabla 2.1: Correspondencia escala de wavelets-contenido de frecuencia de señales.

baja escala a:
$\Rightarrow$ wavelet comprimido
$\Rightarrow$ características que cambian rápidamente
$\Rightarrow$ alta frecuencia ω
alta escala a:
$\Rightarrow$ wavelet estirado
$\Rightarrow$ características que cambian lentamente
$\Rightarrow$ baja frecuencia ω

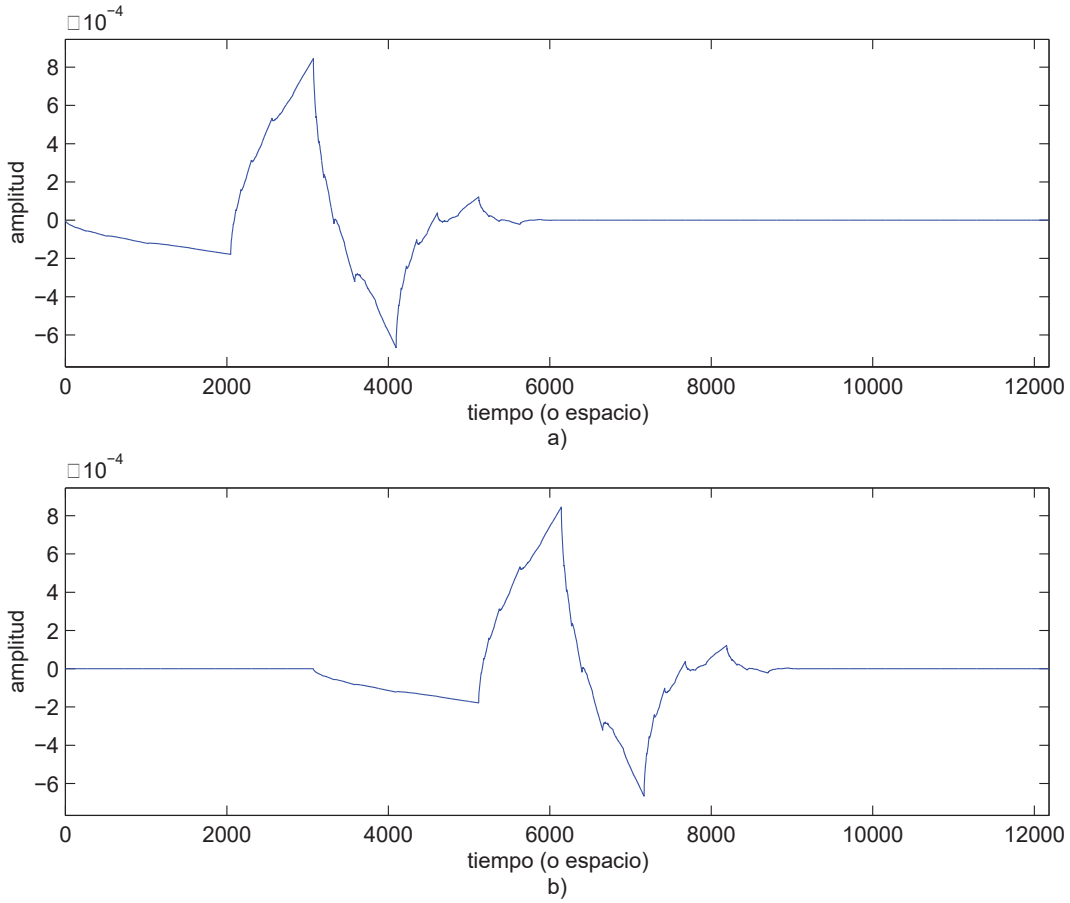

Figura 2.9: Traslación del wavelet madre.

$f\,(t-k)$.

La Figura 2.9a) ilustra la aparición anticipada o retrasada en 2.9b) del Wavelet que se usa como función base en el análisis.

2.3.3. Transformada Wavelet Discreta

Cuando se realiza un análisis de datos mediante la Transformada Wavelet Continua se tiene la posibilidad de calcular coeficientes Wavelet en todas las escalas y posiciones deseadas como la señal bajo estudio permita. Naturalmente, esto implica la inversión de un esfuerzo considerable y la generación de grandes cantidades de datos. Es por ello que surge la necesidad de practicar el análisis de una señal considerando únicamente unas cuantas

escalas y posiciones selectas.

Cuando en el análisis se eligen escalas y posiciones *diádicas*, es decir, que sean potencias de dos, el resultado se obtiene mucho más eficientemente que antes y es casi tan preciso. Un análisis con esta característica se puede obtener de la aplicación de la Transformada Wavelet Discreta (DWT, Discrete Wavelet Transform).

El algoritmo de Mallat [1] [Meyer93, Strang97], que es de hecho un esquema clásico conocido en el ámbito del procesamiento de señales como un codificador en sub-bandas de doble canal [Strang97], constituye un mecanismo eficiente para la implantación del esquema de la Transformada Wavelet Discreta usando filtros. Este algoritmo de filtrado es en efecto un esquema de Transformada Rápida con Wavelets.

Matemáticamente, el conjunto de funciones base formadas por el escalado y la traslación del Wavelet madre $\Phi(x)$ tiene la forma de la ecuación 2.11.

$$\Phi_{(s,l)}(x) = 2^{-\frac{s}{2}} \Phi\left(2^{-s}x - l\right) \tag{2.11}$$

Donde la escala s indica la dilatación del Wavelet madre, y la localización l da su posición.

Para abarcar el dominio de la señal a estudiar en diferentes resoluciones, el Wavelet madre se utiliza en una ecuación de escalamiento con la forma de la ecuación 2.12.

$$W(x) = \sum_{k=-1}^{N-2} (-1)^k c_{k+1} \Phi(2x + k) \tag{2.12}$$

Donde $W(x)$ es la función de escalado para la función madre Φ, y c_k son los coeficientes de *definición* del Wavelet los que no deben ser confundidos con los Coeficientes Wavelet que *resultan* tras la transformación.

[1] Stéphane Georges Mallat. Matemático Francés. Obtuvo su Ph.D. por la Universidad de Pennsylvania en 1988 disertando acerca de Representaciones Multiresolución y Wavelets. Mallat hizo algunas contribuciones fundamentales al desarrollo de la teoría de Wavelets y sus aplicaciones en la caracterización de los transitorios en sonidos y en imágenes a finales de los 1980s y principios de los 1990s. También ha hecho trabajo en matemáticas aplicadas, procesamiento de señales, síntesis de música y segmentación de imágenes. Específicamente, colaboró con Yves Meyer para desarrollar la construcción del Análisis Multiresolución (MRA, Multiresolution Analysis) para Wavelets con soporte compacto, lo cual volvió práctica la implementación de Wavelets para aplicaciones de ingeniería al demostrar la equivalencia de las bases Wavelet y los filtros de espejo conjugados utilizados en bancos de filtros multitasa en procesamiento de señales. También desarrolló (junto a Sifen Zhong) el Método de Máximo del Módulo de la Transformada Wavelet para caracterización de imágenes, un método que usa los máximos locales de los coeficientes Wavelet en varias escalas para reconstruir imágenes.

2.3.3.1. Aproximaciones y detalles

Es de utilidad pensar en los coeficientes de definición del Wavelet $\{c_0, \ldots, c_n\}$ como en los coeficientes de un filtro, los cuales se ubican en una matriz de transformación la cual constituye el núcleo en un mecanismo que se aplica a los datos en la señal bajo análisis. Los coeficientes se ordenan usando dos patrones dominantes, uno que trabaja como un filtro de suavizado o promediador móvil para obtener la *"aproximación"*, y un patrón que actúa como un filtro diferenciador móvil para obtener el *"detalle"* de la información en los datos. Este par de ordenamientos en los coeficientes se denomina *filtro de espejo en cuadratura* [Strang97]. La matriz de coeficientes opera sobre los datos siguiendo un algoritmo jerárquico, que a menudo se denomina *algoritmo piramidal* [Strang97]. Los coeficientes del Wavelet se organizan de manera que los renglones impares en la matriz contengan un ordenamiento de coeficientes que actúan como el filtro suavizador, y los renglones pares de la matriz contengan un ordenamiento de coeficientes con signos diferentes que actúen para obtener el detalle en los datos. El proceso considera primero el total de los datos para ser suavizados y al término de esta fase quedan reducidos en número a la mitad del total original. Luego el procedimiento se aplica otra vez para suavizar esta nueva cantidad de datos y hecho esto, los datos nuevamente se han reducido en número a la mitad del total inmediato anterior, lo que equivale a pensar que se cuenta aproximadamente con información que en volumen representa la cuarta parte del volumen de datos al inicio del estudio. Este proceso continúa hasta que el suavizado de los datos que resten sea trivial. En otras palabras, cada aplicación de la matriz de transformación extrae una mejor resolución de los datos, mientras continúa suavizando los restantes. La salida de la Transformada Wavelet Discreta consiste de los últimos datos resultantes del suavizado final y de todas las componentes de "detalles" acumuladas.

En muchas señales, el contenido de baja frecuencia es la parte más importante. Por ejemplo, en el caso de la voz humana si se retiran los componentes de más alta frecuencia aún cuando la voz suena distinto, todavía se puede identificar lo que se ha dicho. En contraparte, si de una señal de voz humana se remueven suficientes componentes de baja frecuencia, lo único que se escuchará es "siseo" ininteligible. Es por esta razón que en el análisis con

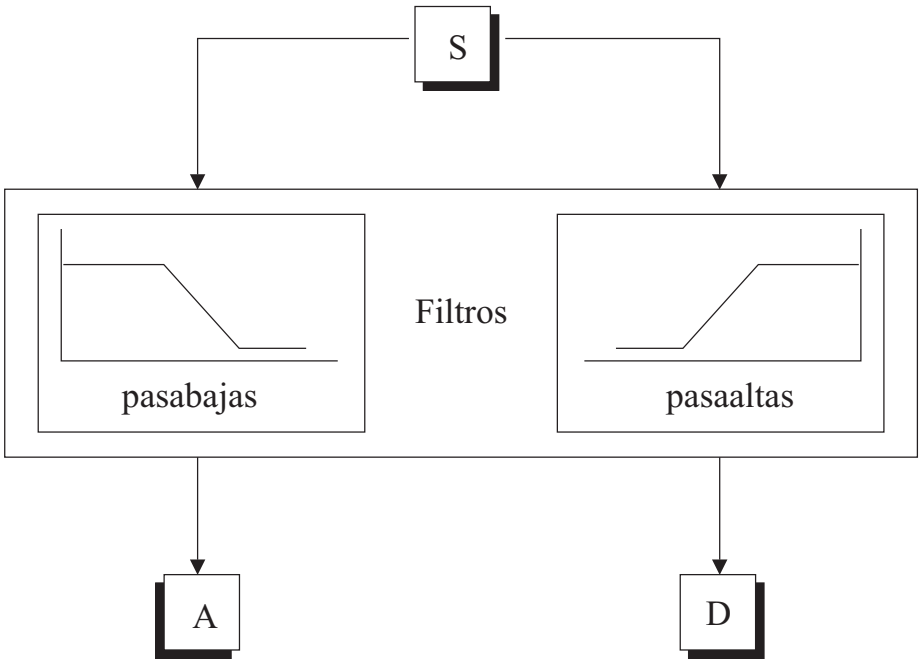

Figura 2.10: Derivación de aproximaciones A y detalles D mediante el proceso de filtrado de la señal S.

Wavelets se habla generalmente de aproximaciones y detalles. Las aproximaciones son los componentes de alta escala o de baja frecuencia en la señal. Los detalles son los componentes de baja escala o de alta frecuencia en los datos [Misiti96].

El proceso de filtrado en su nivel más básico se ilustra en la Figura 2.10.

La señal bajo estudio S hace pasar a través de dos filtros complementarios y emerge como dos señales A y D que constituyen la aproximación y el detalle respectivamente. Por desgracia, si esta operación se practica sobre una señal digital real, se termina acumulando el doble de datos de la señal original como se percibe en la Figura 2.11.

Para corregir este problema se submuestrea la información de salida de cada uno de los filtros. Esto equivale a simplemente desechar cada segunda muestra en los datos a la salida de cada uno de los filtros. Este proceso es el que produce los Coeficientes de la DWT cA y cD que constituyen los coeficientes de Aproximación y de Detalle respectivamente, para los que la acumulación de datos resultantes constituye el mismo volumen de información que la señal original. Una ilustración de este mecanismo se puede observar en la Figura 2.12.

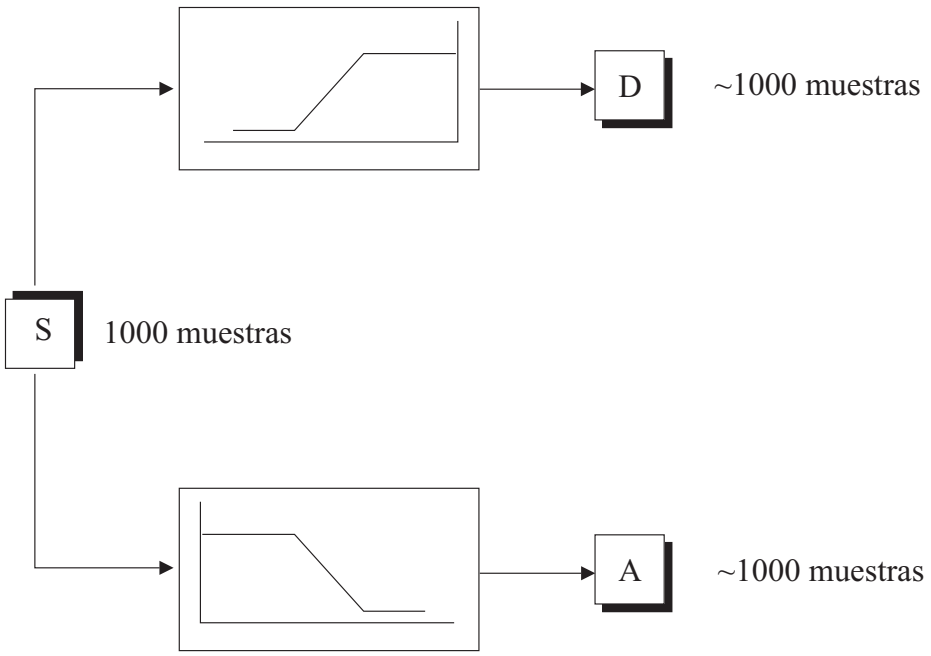

Figura 2.11: Filtrado de una señal S sin submuestreo para la obtención de A y D.

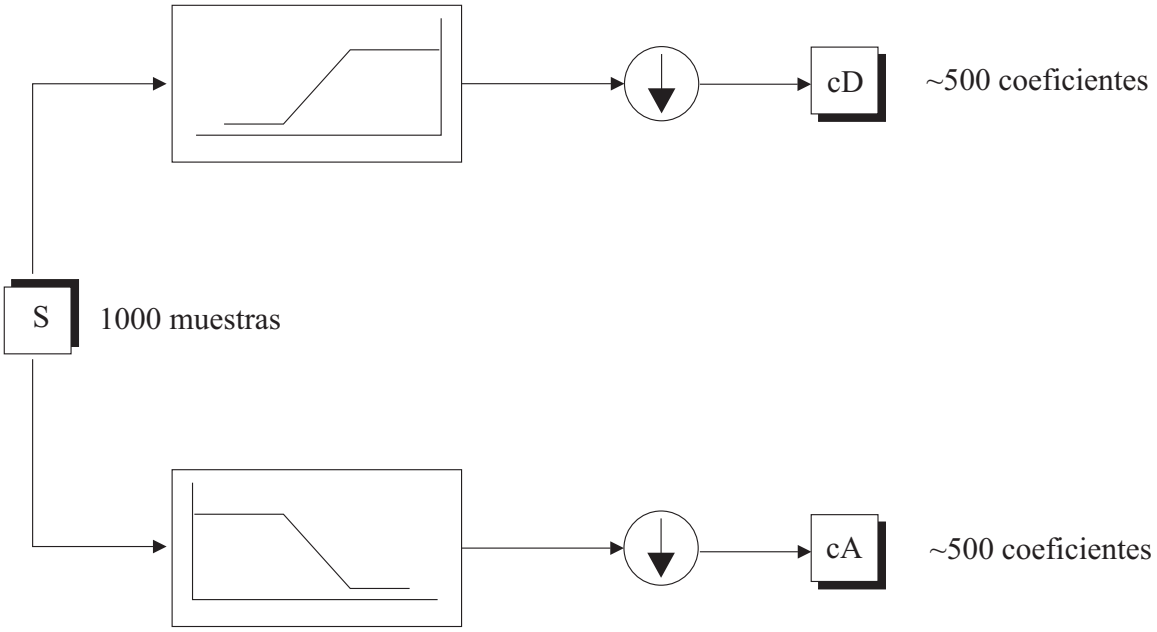

Figura 2.12: Obtención de los Coeficientes Wavelet de Aproximación cA y Detalle cD mediante el filtrado con submuestreo de una señal S.

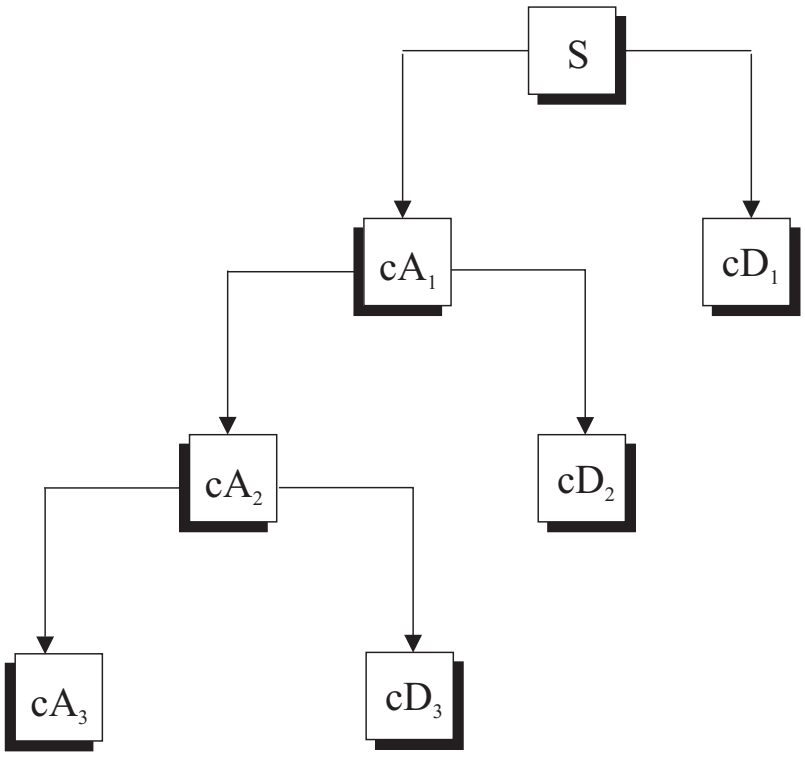

Figura 2.13: Descomposición multinivel de una señal *S*.

El proceso de filtrado y submuestreado realiza la descomposición de la señal S en un nivel o escala.

2.3.3.2. Descomposición multinivel

Al haberse efectuado el filtrado de los datos en una primera ocasión, el proceso de descomposición puede llevarse a cabo una vez más, considerando los nuevos coeficientes de Aproximación *cA* como la entrada para la fase de filtrado y submuestreo que se está empleando. De este modo, la señal original se descompone para obtener cada vez una nueva representación con menor resolución. Los resultados de varias repeticiones del proceso dan origen a una representación jerárquica en la que los coeficientes de Aproximación y de Detalle se ordenan en una estructura de árbol, como se puede apreciar en la Figura 2.13.

La Figura 2.14 ilustra cuatro niveles de descomposición de la palabra *"uno"* pronunciada por un usuario. Aún cuando en teoría la descomposición podría seguirse efectuando

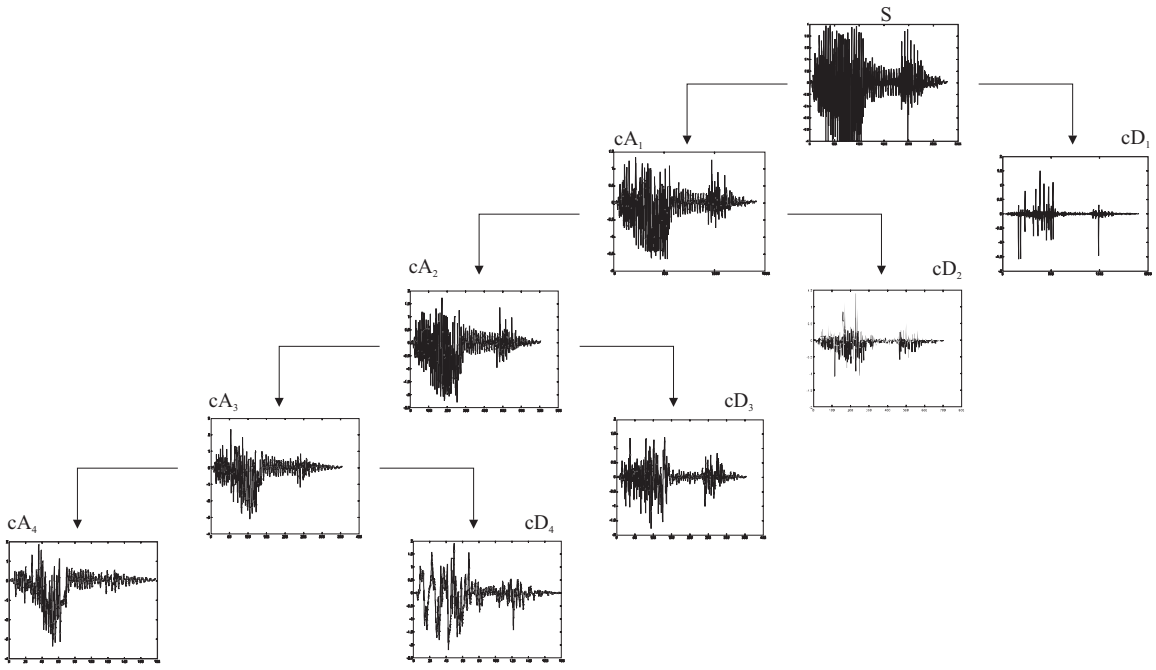

Figura 2.14: Descomposición multinivel de una elocución de la palabra *"uno"*.

en los siguientes niveles indefinidamente, la realidad es que el máximo nivel permisible es aquel en el que hay un sólo coeficiente de Aproximación y uno de Detalle.

2.3.4. El Wavelet de Haar

El Wavelet de Haar es el primero en ser descubierto a principios del siglo 20 [Meyer93] por Alfrèd Haar [2] y también es por mucho el más simple, aún cuando su utilidad para fines prácticos es limitada dado el hecho de que no es continuamente diferenciable.

Matemáticamente, el Wavelet de Haar se define mediante la ecuación 2.13.

$$\Psi(t) = \begin{cases} 1 \,\forall\, t \in \left[0, \frac{1}{2}\right) \\ -1 \,\forall\, t \in \left[\frac{1}{2}, 1\right] \\ 0 \,\forall\, t \notin (0, 1) \end{cases} \tag{2.13}$$

[2] Alfrèd Haar : (11 de Octubre de 1885 - 16 de Marzo de 1933). Matemático Húngaro. En 1904 comenzó sus estudios en la Universidad de Göttingen. Su doctorado fue supervisado por David Hilbert. La medición de Haar, el Wavelet de Haar y la transformada Haar se nombran de esas maneras en su honor. Junto a Frigyes Riesz hizo de la Universidad de Szeged un centro de matemáticas. También fundó la revista Acta Scientiarum Mathematicarum junto a Riesz.

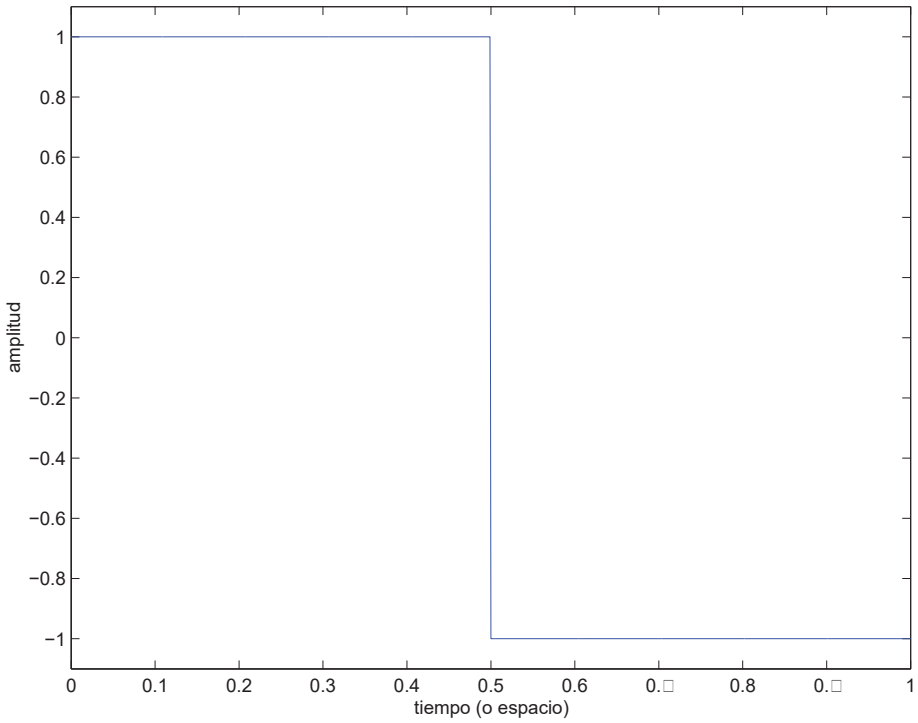

Figura 2.15: Wavelet de Haar.

Los valores de los coeficientes que definen el filtro pasabajas de descomposición correspondiente al wavelet de Haar se muestran en la ecuación 2.14, y estos se emplean en el cálculo de los coeficientes wavelet de la información en la señal de acuerdo con el algoritmo que se presenta en la sección 2.4.1.

$$[L] = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right] \tag{2.14}$$

Una ilustración de la simplicidad en la forma de onda del Wavelet de Haar se presenta en la Figura 2.15. Algunos autores consideran este Wavelet como el de Daubechies de primer orden, y por tanto, se le identifica de manera abreviada como el Wavelet db01 [Misiti96].

2.3.5. Los Wavelets de Daubechies

Uno de los descubrimientos que en años recientes Ingrid Daubechies [3] realizó en el campo de los Wavelets a partir del trabajo que Mallat realizó a mediados de los 1980's, es el hecho de que las funciones base para estos se pueden obtener partiendo de los valores de los coeficientes que definen filtros discretos [Meyer93]. Con ello se abrió la posibilidad de conformar familias de funciones base localizadas en el espacio y continuamente diferenciables. La forma de onda de algunas versiones del Wavelet de Daubechies de orden 2 se presentó en las Figuras 2.8 y 2.9.

Los valores de los coeficientes que definen el filtro pasabajas L de descomposición correspondiente al Wavelet de Daubechies de orden 2 pueden observarse en la ecuación 2.15, y estos se emplean en el cálculo de los coeficientes wavelet de la información en la señal de acuerdo con el algoritmo que se presenta en la sección 2.4.2.

$$[L] = \left[\frac{1+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{1-\sqrt{3}}{4\sqrt{2}} \right] \tag{2.15}$$

2.4. Implantación de un algoritmo de DWT

El núcleo en la implantación de un algoritmo para calcular Coeficientes Wavelet mediante una Transformada Discreta, consiste en una acumulación de productos de los coeficientes que definen el filtro pasabajas con las muestras en la señal que se analiza para toda la duración de esta.

2.4.1. DWT basada en el Wavelet de Haar

Si se emplea el Wavelet de Haar, el primer coeficiente Wavelet de Aproximación se obtiene de la suma de: el producto del coeficiente L_0 del filtro con la primera muestra en la señal, y el producto del coeficiente L_1 del filtro con la segunda muestra en la señal.

[3] Ingrid Daubechies (17 de Agosto de 1954 -). Matemática y Física Belga nacida en la ciudad de Houthalen. Daubechies completó sus estudios de física en la Universidad de Vrije in 1975. Obtuvo su Ph.D. en física teórica en 1980, y continuó su carrera de investigación en esa institución hasta 1987. Sus trabajos mejor conocidos son: el Wavelet ortogonal de Daubechies y el Wavelet biortogonal CDF. Un Wavelet de esta familia de Wavelets se usa en la actualidad en el estándar JPEG 2000.

El segundo coeficiente Wavelet de Aproximación se obtiene sumando: el producto del coeficiente L_0 con la tercera muestra en la señal y el producto del coeficiente L_1 con la cuarta muestra en la señal. El proceso se itera para el total de muestras en la señal. Con ello se obtienen todos los coeficientes de Aproximación para un nivel.

El algoritmo para el cálculo de los coeficientes de Aproximación y de Detalle en un nivel determinado implantando la Transformada Discreta con Wavelets empleando el Wavelet de Haar, se ilustra en el Algoritmo 1.

Algoritmo 1 DWT usando el wavelet de Haar

```
TRANSFORMADAWAVELETHAAR(señal, nivel)
1  coeficientesACalcular ← longitudDeSeñal/2
2  l0 ← l1 ← 1/√2
3  para i = 0 hasta i < nivel
4    para j = 0 hasta j < coeficientesACalcular
5      aproximacion[i][j] ← (señal[j * 2] * l0 + señal[j * 2 + 1] * l1)
6      detalle[i][j] ← (señal[j * 2] * −l0 + señal[j * 2 + 1] * l1)
7    señal ← aproximacion
8    coeficientesACalcular ← coeficientesACalcular/2
```

2.4.2. DWT basada en el Wavelet de Daubechies de orden 2

Si se emplea el Wavelet de Daubechies de orden 2, cada coeficiente Wavelet de Aproximación se obtiene con la suma de 4 productos entre los coeficientes que determinan el filtro pasabajas del Wavelet con cuatro muestras en la señal. El segundo coeficiente Wavelet de Aproximación se obtiene de la sumatoria de los productos: L_0 con la primera muestra en la señal, L_1 con la segunda muestra, L_2 con la tercera muestra y L_3 con la cuarta muestra. El tercer coeficiente Wavelet de Aproximación se obtiene de la sumatoria de productos: L_0 con la tercera muestra en la señal, L_1 con la cuarta muestra, L_2 con la

quinta muestra y L_3 con la sexta muestra. El proceso se itera para el total de muestras en la señal. Con ello se obtienen todos los coeficientes de Aproximación para un nivel. Esta metodología produce un problema en los extremos de la señal, donde se cuenta con más coeficientes de filtro por multiplicar que muestras en la señal, de este modo, para el cálculo del primero y el último de los coeficientes Wavelet se consideran los dos primeros datos y los dos últimos respectivamente, como reflejados en un espejo al considerar el inicio y el final de la señal. Así entonces, el primer coeficiente Wavelet de Aproximación se calcula como la sumatoria de los productos: L_0 con la segunda muestra en la señal, L_1 con la primera muestra, L_2 con la primera muestra y L_3 con la segunda muestra. De manera análoga, el último coeficiente Wavelet de Aproximación se calcula con la suma de los productos: L_0 con la penúltima muestra en la señal, L_1 con la última muestra, L_2 con la última muestra y L_3 con la penúltima muestra.

El algoritmo para el cálculo de los coeficientes de Aproximación en un nivel determinado implantando la Transformada Discreta con Wavelets empleando el Wavelet de Daubechies de orden 2, se ilustra en el Algoritmo 2.

2.5. Caracterización de la señal de voz con Wavelets

El proceso de descomposición de una señal utilizando Wavelets permite obtener una nueva representación de los datos originales en cada nuevo nivel del análisis. Como se mencionó antes, la señal de la voz humana porta esencialmente en sus contenidos de baja frecuencia la información acerca de qué palabra se está pronunciando, y además los coeficientes Wavelet de Aproximación son el resultado de hacer pasar la señal a través de un filtro pasabajas.

Es por ello que se puede descomponer la señal hasta cierto nivel conveniente en el que la descomposición retenga aún suficiente información significativa de baja frecuencia, y considerar los datos originales como caracterizados por los coeficientes Wavelet de Aproximación correspondientes a ese nivel.

Es importante recordar que el efecto de descomponer la señal para obtener los coeficientes Wavelet en un nuevo nivel, produce una representación más suavizada de los

Algoritmo 2 DWT usando el wavelet de Daubechies de orden 2

```
TRANSFORMADAWAVELETDAUBECHIES(señal, nivel)
 1  coeficientesACalcular ← longitudDeSeñal/2
 2  l0 ← (1 + √3) / (4√2)
 3  l1 ← (3 + √3) / (4√2)
 4  l2 ← (3 − √3) / (4√2)
 5  l3 ← (1 − √3) / (4√2)
 6  para i = 0 hasta i < nivel
 7    aproximacion [i] [0] ← (señal [1] ∗ l3 + señal [0] ∗ l2) +
 8    (señal [0] ∗ l1 + señal [1] ∗ l0)
 9    detalle [i] [0] ← (señal [1] ∗ −l0 + señal [0] ∗ l1) +
10    (señal [0] ∗ −l2 + señal [1] ∗ l3)
11    para j = 0 hasta j < coeficientesACalcular
12      aproximacion [i] [j + 1] ← (señal [j ∗ 2] ∗ l3 + señal [j ∗ 2 + 1] ∗ l2) +
13      (señal [j ∗ 2 + 2] ∗ l1 + señal [j ∗ 2 + 3] ∗ l0)
14      detalle [i] [j + 1] ← (señal [j ∗ 2] ∗ −l0 + señal [j ∗ 2 + 1] ∗ l1) +
15      (señal [j ∗ 2 + 2] ∗ −l2 + señal [j ∗ 2 + 3] ∗ l3)
16    aproximacion [i] [coeficientesACalcular] ← (señal [ultimo] ∗ l3) +
17    (señal [penultimo] ∗ l2 + señal [penultimo] ∗ l1 + señal [ultimo] ∗ l0)
18    detalle [i] [coeficientesACalcular] ← (señal [ultimo] ∗ −l0) +
19    (señal [penultimo] ∗ l1 + señal [penultimo] ∗ −l2 + señal [ultimo] ∗ l3)
20    señal ← aproximacion
21    coeficientesACalcular ← coeficientesACalcular/2
```

datos en los coeficientes de Aproximación, de tal manera que si se filtra en exceso se corre el riesgo de perder las características esenciales de baja frecuencia que distinguen una palabra de otra. Por otra parte, al avanzar de un nivel a otro en la descomposición, el volumen de datos que debe manejarse para caracterizar una señal se reduce a la mitad, de manera que debe buscarse el mejor equilibrio entre la cantidad de coeficientes de Aproximación que caracterizan una señal, y la calidad desde el punto de vista del reconocimiento de palabras que ofrece la representación en el nivel correspondiente [Janer96, Gillemain96, Wesfreid93].

En este punto se ha completado la delicada tarea de caracterizar las señales acústicas de una manera conveniente y los resultados deberán procesarse en la etapa responsable de la emisión del juicio respecto de la identidad de la información contenida en la señal. La detección de esta "marca" de pertenencia brindará la posibilidad de clasificar las diversas instancias de elocución de palabras que sean analizadas bajo la clase que les corresponda a cada una. Esta problemática será abordada en el Capítulo 3.

Capítulo 3

Similitud y clasificación de patrones

La tarea de clasificación constituye un proceso en el cual han de formarse grupos de entidades que comparten características similares. Cada vez que una nueva entidad se somete al proceso, habrán de detectarse preponderantemente en ella las características distintivas en los miembros de alguna de las clases previamente formadas, y entonces se le anexará a esa clase en particular, evitando asociarle con una clase donde los elementos no demuestren poseer tal característica en absoluto, o exhiban esa característica pero en una medida relativamente menor que otras. La similitud o semejanza debe poder estimarse cualitativamente a través de un sistema de medición adecuado, el cual permita emitir un juicio lo más certero posible respecto del grado de pertenencia de un elemento a uno u otro conjunto mutuamente disjuntos. Este capítulo presenta algunas de las técnicas popularmente usadas en la estimación de las medidas de similaridad entre patrones, y para el agrupamiento o formación de "grumos" a partir de la acumulación de entidades con relativa cercanía entre sí.

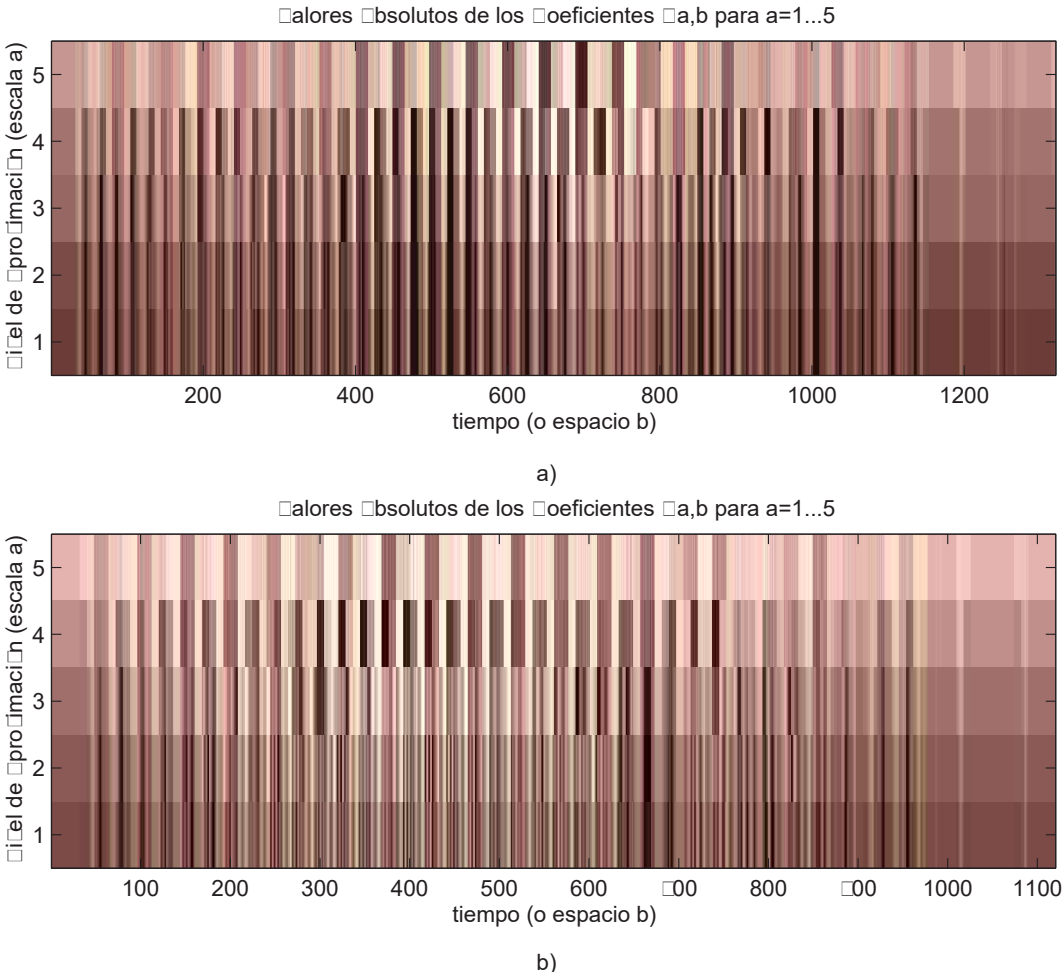

Figura 3.1: Escalagramas de la elocución de dos palabras distintas: *"uno"* (arriba) y *"cero"* (abajo).

3.1. Similitud entre dos patrones

La Figura 3.1 muestra dos escalagramas que contienen una representación gráfica para los valores absolutos de los coeficientes de aproximación en las primeras 5 escalas de descomposición, calculados con la Transformada Wavelet Discreta basada en el wavelet de Haar en la elocución de dos palabras distintas ("uno" en la Figura 3.1a) y "cero" en la Figura 3.1b)). Al examinar cuidadosamente este gráfico, se pueden detectar diferencias significativas entre ambos escalagramas, que inducen a conjeturar que corresponden a palabras diferentes.

Más aún, una misma palabra puede pronunciarse de distinta manera en diferentes

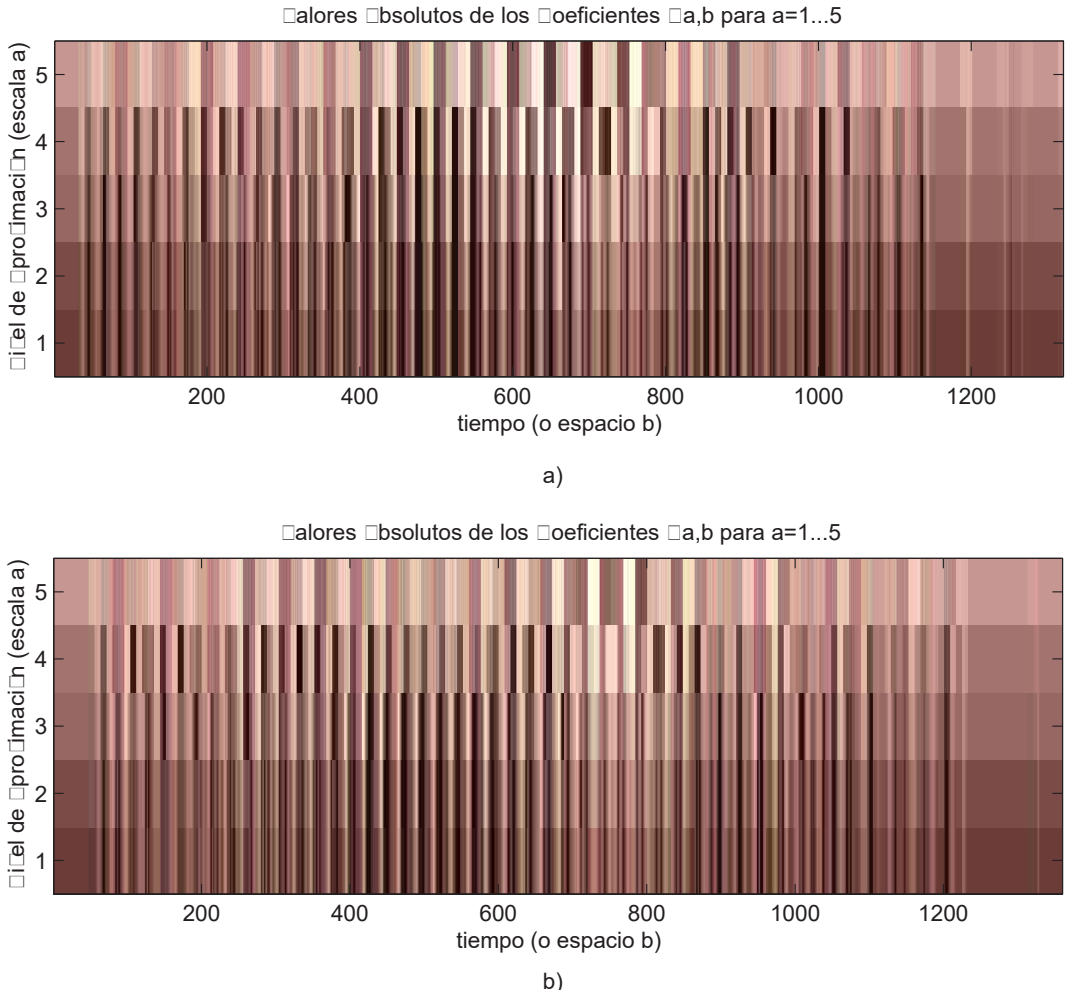

Figura 3.2: Escalagramas de dos elocuciones de la misma palabra: *"uno"*.

ocasiones por una amplia gama de razones, algunas de las cuales se mencionaron antes en la sección 1.2.2.

La Figura 3.2 muestra dos escalagramas que contienen una representación gráfica para los valores absolutos de los coeficientes de aproximación en las primeras 5 escalas de descomposición, calculados con la Transformada Wavelet Discreta basada en el wavelet de Haar en dos elocuciones de la palabra "uno". Al examinar cuidadosamente este gráfico, se pueden detectar amplias similitudes entre ambos escalagramas, aunque no son exactamente iguales.

En un sistema de reconocimiento automático de voz se prefiere la comparación

de los patrones descritos por las representaciones simplificadas de los datos que ofrece la etapa de caracterización de las señales, la cual se mencionó en el Capítulo 2. Es decir, en el esquema que se propone, esta fase del problema se ataca comparando el vector de Coeficientes Wavelet de Aproximación de una señal en cierto nivel adecuado, contra el vector de Coeficientes Wavelet de Aproximación de otra señal en el mismo nivel de su contraparte.

Los términos "cercanas" y "lejanas" tienen una connotación intrínseca de distancia geométrica, bien proximidad o separación, lo cual puede ser medido y asociado con una cifra que elimine toda posibilidad de ambigüedad al estimar la similitud entre dos patrones.

Este capítulo está dedicado a la descripción de algunos de los mecanismos que pueden emplearse para medir la distancia existente entre los patrones que describen dos señales para efectos del reconocimiento automático de voz, enfatizando en una estrategia que se enfoca especialmente a la atenuación del efecto de alargar o acortar la duración de una palabra por haberse pronunciado más lentamente en una segunda ocasión, o por haberse pronunciado más aprisa en una segunda ocasión.

3.2. Distancia entre dos vectores de características de la señal de voz

Cualquier metodología que se utilice para la medición de la distancia existente entre dos vectores de Coeficientes Wavelet de Aproximación x e y que representen la pronunciación de dos palabras debe poseer las siguientes propiedades:

a) La distancia entre el vector x y el vector y es un número real positivo mayor a cero (definición positiva), es decir:

$$d(x, y) > 0 \tag{3.1}$$

b) La distancia entre el vector x y el mismo vector x es cero, o sea:

$$d(x, x) = 0 \tag{3.2}$$

c) La distancia entre el vector x y el vector y, tiene el mismo valor que la distancia entre el vector y el vector x (condición de simetría), matemáticamente:

$$d(x, y) = d(y, x) \tag{3.3}$$

d) La distancia entre el vector x y el vector y, es menor o a lo sumo igual que la suma de la distancia entre los vectores x y z y la distancia entre los vectores z y y (condición de desigualdad triangular), es decir:

$$d(x, y) \leq d(x, z) + d(z, y) \tag{3.4}$$

3.2.1. Distancias L_p

La distancia que separa dos vectores x e y que almacenan cada uno un patrón de una señal se puede calcular con la ecuación 3.5.

$$d_{L_p}(x, y) = \left[\sum_{i=1}^{N} |x_i - y_i|^p \right]^{\frac{1}{p}} \tag{3.5}$$

donde i es el i-ésimo elemento en cada vector y N es la dimensión de los vectores x e y. Se trata de una familia de métricas de agrupadas como distancias de Minkowsky, entre las que destacan las que a continuación se enuncian.

3.2.1.1. Distancia L_1

Cuando p en 3.5 es igual a la unidad, se está calculando la distancia L_1 de acuerdo con la ecuación 3.6.

$$d_{L_1}(x, y) = \sum_{i} |x_i - y_i| \tag{3.6}$$

Esta métrica se denomina como la distancia rectilínea, distancia o longitud de Manhattan, distancia de cuadra de ciudad, entre otros nombres.

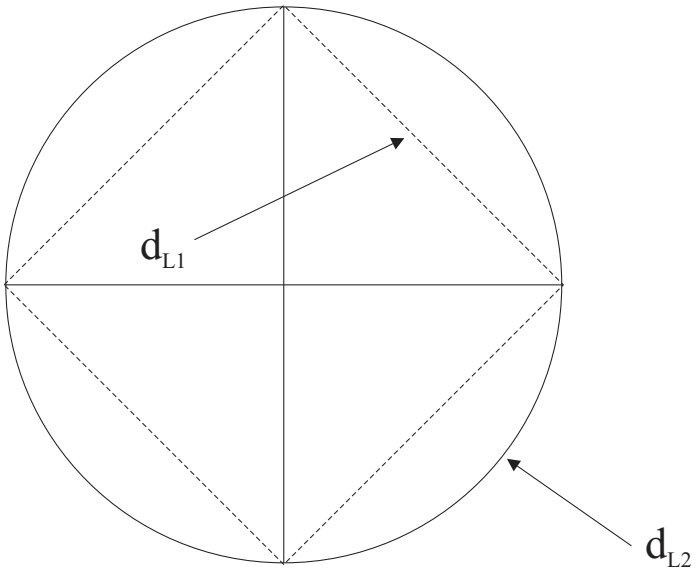

Figura 3.3: Interpretación geométrica de algunas de las distancias L_p.

3.2.1.2. Distancia euclidiana o L_2

Si p en la ecuación 3.5 es igual a 2, se está calculando la distancia L_2, que se conoce también con el nombre de distancia euclidiana de acuerdo con la ecuación 3.7.

$$d_{L_2}(x, y) = \sqrt{\sum_i |x_i - y_i|^2} \tag{3.7}$$

La Figura 3.3 muestra una interpretación geométrica que puede darse a las distancias L_p. En L_1 los puntos equidistantes del centro u origen se encuentran localizados sobre el perímetro de un cuadrado, mientras que en L_2 puntos que distan la misma magnitud del punto central se hallan ubicados sobre el perímetro de una circunferencia.

Existen otras clases de distancias entre patrones tal como la de Mahalanobis, la cual a diferencia de la distancia euclidiana, se basa en las correlaciones entre variables y es invariante a la escala de las mediciones. En ocasiones también se emplean medidas de similitud que no son distancias, tales como la distancia coseno, la cual es utilizada en recuperación de información donde un documento se representa como un vector de pesos.

3.3. Alineación temporal

Para estimar la distancia entre dos patrones, podría pensarse que es suficiente la comparación de los elementos en las posiciones correspondientes del vector de características de una señal y otra, desde el inicio hasta el final, es decir, considerar diferencias locales coeficiente a coeficiente. Esto funciona adecuadamente si ambos vectores tienen la misma longitud, condición que en el ámbito del reconocimiento de palabras, en la mayoría de las ocasiones no se cumplirá. Para tener plena certeza de contar con dos vectores de características de igual longitud, se haría necesario que las señales originales tuviesen el mismo número de muestras como consecuencia de haberse empleado precisamente el mismo tiempo en la pronunciación de ambas palabras. Este es un escenario muy difícil de obtener para el caso de la pronunciación en dos distintas ocasiones de una sola palabra, y aún más difícil de conseguir para el caso de la pronunciación de dos palabras con distinto número de fonemas.

Con el fin de efectuar una normalización en tiempo, se puede "alargar" el patrón más corto insertando nuevos datos interpolados en base a los existentes, hasta que el número de Coeficientes Wavelet de Aproximación del patrón más corto iguale al número de Coeficientes Wavelet de Aproximación del patrón más largo.

También es es posible el "acortamiento" del patrón más largo eliminando algunos de los datos que contiene, hasta que el número de Coeficientes Wavelet de Aproximación del patrón más largo iguale al número de Coeficientes Wavelet de Aproximación del patrón más corto.

El primer problema en el hipotético caso de que se decidiera seguir un esquema como este es: ¿se utiliza el criterio de alargar el patrón más corto, o el criterio de acortar el patrón más largo?

Una vez resuelto el primer problema el siguiente consiste en decidir: si se alarga el patrón más corto ¿se incrustan nuevos datos al inicio, al final, o en el interior del patrón?; si se acorta el patrón más largo ¿se eliminan datos al inicio, al final o en el interior del patrón?

Es evidente que para implantar un esquema de esta naturaleza se hace necesario tomar algunas decisiones que no siempre tendrán un sustento objetivo. Por ello, quizá lo más sensato sea conservar los vectores que almacenan las características de las palabras

inalterados para efectuar su comparación.

3.3.1. Doblado Dinámico en Tiempo

Un algoritmo que permite calcular una distancia global entre dos vectores de características de una señal, como lo pueden ser los vectores de Coeficientes Wavelet de Aproximación, es el Doblado Dinámico en Tiempo (DTW, Dynamic Time Warping). Este es un método que permite alinear temporalmente dos vectores que no tienen la misma longitud [Myers81, Mariani89].

El objeto del uso de Doblado Dinámico en Tiempo es el de registrar[1] el primer elemento de uno de los vectores con el primer elemento del otro vector entre sí. Del mismo modo, se efectúa el registro del último elemento del primer vector con el último elemento del segundo vector. Entre el inicio y fin de los vectores, el Doblado Dinámico en Tiempo proporciona una decisión óptima de cuál debe ser el orden de registro de los elementos en el interior de cada uno de los vectores. Cada vez que se da un paso en la comparación de elementos en los vectores, se acumula una medida de distancia local con lo que al llegar a la comparación de los últimos elementos en los vectores se cuenta con una medida de distancia global entre los vectores.

El problema del alineamiento temporal se ilustra en la Figura 3.4, en la que la duración del primero de los patrones se extiende desde la izquierda hacia la derecha y la duración del segundo patrón se extiende desde abajo hacia arriba. La línea en el interior de la "retícula" o *matriz de tiempo* simboliza el trayecto óptimo de alineación entre ambos patrones, es decir, cuál elemento del primer patrón debe ser comparado con cuál elemento del segundo patrón, de manera que la distancia entre ambos elementos (su discrepancia) sea la mínima posible. Cada una de las distancias locales entre elementos individuales en los vectores es acumulada a lo largo del recorrido sobre ambos patrones, con lo que finalmente al hacer la comparación entre el último dato en uno y otro vector, de la acumulación se encuentra una distancia global.

El dejar que transcurra el tiempo para uno de los patrones y no para el otro,

[1] Se hace uso del vocablo con la connotación aplicable al terreno de la imprenta, en el que se refiere a la *correspondencia igual de las planas de un pliego impreso con las del dorso.*

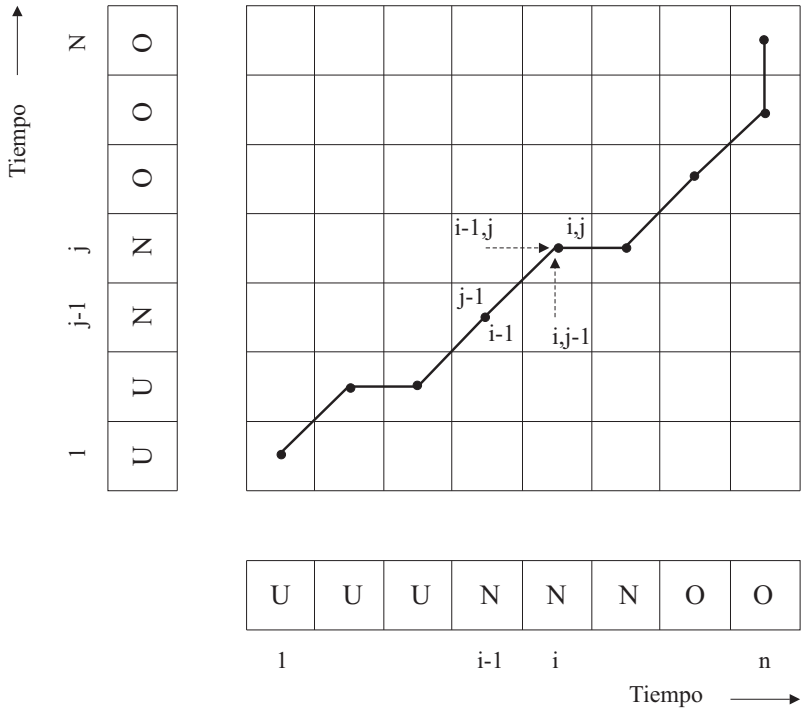

Figura 3.4: Matriz de alineación de patrones con Doblado Dinámico en Tiempo.

equivale a la eliminación de marcos en la señal, lo cual puede ser susceptible de penalización, la que se modela como una "función de peso" que puede incrementar la magnitud de la distancia local cuando se presenta esta "falta de avance" en uno de los dos patrones.

Resulta obvio que el método de Doblado Dinámico en Tiempo no explora todos los trayectos posibles en la retícula de tiempo, sino que es necesario imponer ciertas restricciones al proceso de tomar la decisión de preferir un elemento sobre otro para usarlo en la comparación. Estas sencillas restricciones son:

1. Los trayectos de alineación de patrones no pueden retroceder en el tiempo.
2. Cada elemento en cada uno de los patrones debe ser usado en el trayecto de alineación.
3. Las medidas de distancias locales se combinan a través de una sumatoria para arrojar una distancia global.

La Figura 3.5 ilustra las tres posibilidades de movimiento del punto más reciente al siguiente en el proceso de alineación temporal con Programación Dinámica aplicando

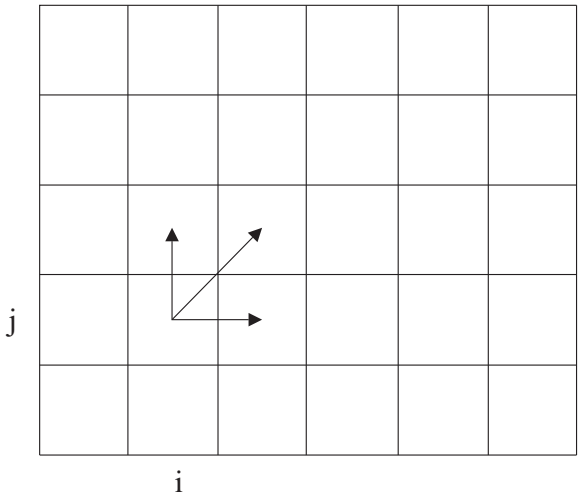

Figura 3.5: Tres movimientos posibles de un punto al siguiente en el trayecto de alineación temporal empleando restricciones locales simétricas de primer orden.

restricciones locales de primer orden, las cuales consideran como origen del siguiente movimiento a los tres puntos previos inmediatos. Estas restricciones locales de continuidad se describen generalmente especificando el trayecto completo en términos de trayectos locales simples, los cuales pueden ser reunidos entre sí para formar trayectos cada vez más largos.

Es razonable pensar también en otras restricciones locales además de las anteriormente citadas, las cuales permitan el avance a un punto subsecuente desde puntos previos que se encuentran más alejados que los tres próximos inmediatos [Myers81]. Por ejemplo, la Figura 3.6 muestra restricciones locales de tipos I a III y de Itakura que permiten arribar al punto (n,m).

Si se considera un punto dentro de las celdas en la matriz de tiempo, digamos la que tiene los índices (i, j), el punto previo en el trayecto que se ha recorrido debió ser bien $(i-1, j-1), (i-1, j)$, o $(i, j-1)$ de acuerdo con la restricción simétrica de primer orden, como se aprecia en la Figura 3.4. La idea central del Doblado Dinámico en Tiempo de acuerdo con la restricción local simétrica de primer orden también conocida como de Itakura, es que se debe avanzar hacia el punto en (i, j), desde el punto previo de los tres posibles, el cual representa la menor distancia al punto en (i, j). Entonces, si $D(i, j)$ es la distancia global hasta el punto (i, j) y $d(i, j)$ expresa la distancia local al punto (i, j) :

$$D(i,j) = m\acute{i}nimo\left[D(i-1,j-1),\ D(i-1,j),\ D(i,j-1)\right] + d(i,j)$$

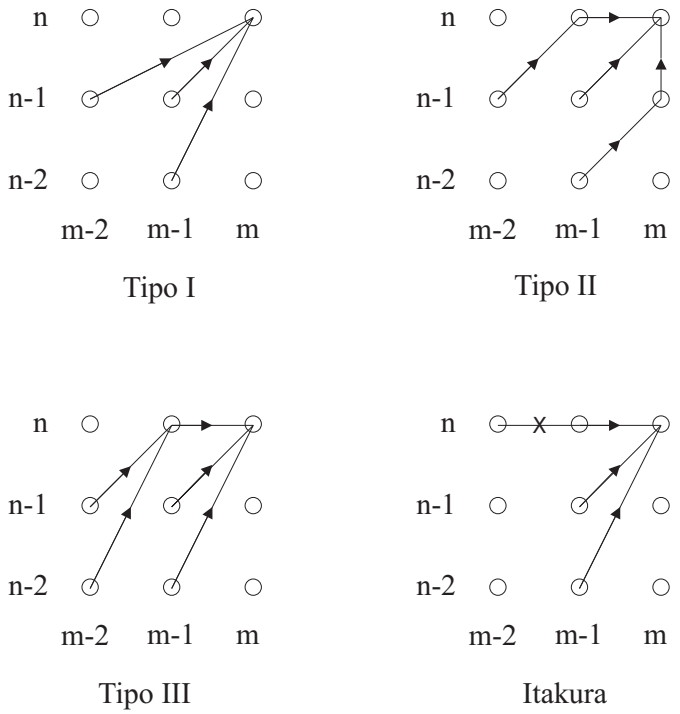

Figura 3.6: Restricciones locales usadas para DTW.

y

$$D(1,1) = d(1,1)$$

La distancia global $D(N,n)$ proporciona una medida general de cuánto se acercan ambos patrones comparados.

Dado que se tiene la condición inicial de que $D(1,1) = d(1,1)$, esta puede constituir la base de un algoritmo recursivo para la implantación del Doblado Dinámico en Tiempo, sin embargo el proceso del cálculo de distancia entre dos vectores aún de longitudes reducidas puede consumir gran cantidad de tiempo, dado lo cual, la implementación se lleva a cabo mediante Programación Dinámica con un algoritmo como el que se muestra en la sección 4.4.

3.4. Clasificación de patrones

Los atributos que caracterizan inequívocamente a un objeto pueden también encontrarse en otro objeto, de manera que se les califique como similares o con cierto parecido si una cantidad suficiente de tales atributos o características se encuentra presente en ambos. A menudo no se tiene certeza de cuáles son precisamente los atributos que deben considerarse al comparar los objetos, así que en la mayoría de los casos, lo mejor que se puede afirmar es si dos objetos pertenecen a una misma clase o a clases diferentes. De esta manera, el reconocer algún objeto como similar a otro es un análogo de la tarea de clasificación o agrupamiento de pares.

En el reconocimiento automático de voz, el patrón de caracterización de una señal recién obtenido debe compararse con otros patrones en su misma clase y en otras clases diferentes previamente almacenados, de manera que el resultado de esta comparación indique en cuál de las clases conocidas debe ubicarse al nuevo elemento. En otras palabras, el problema que se debe resolver ahora es que dado un conjunto de vectores de características de referencia que pertenecen a clases conocidas y un nuevo vector de características de prueba del que se desconoce la clase a la que pertenece, debe indicarse la clase a la que debe unirse este vector de prueba que es la misma clase a la que pertenece el vector de referencia con el que guarda mayor similaridad.

Los patrones de referencia cuyas clases son conocidas para el sistema automático de reconocimiento se almacenan en el diccionario del sistema, que contiene todas las palabras que el sistema puede reconocer.

Las secciones siguientes están dedicadas a hacer mención de un par de estrategias de clasificación de patrones, que se emplean de manera común en la etapa de un sistema de reconocimiento de voz que emite el resultado de la caracterización de una señal de entrada en términos de los símbolos del lenguaje respectivos.

3.5. El vecino más cercano

El método del vecino más cercano es de suma importancia dada la posibilidad de aplicarlo en la solución de una amplia variedad de ámbitos ente los que se encuentran: bases

de datos, minería de datos y, por supuesto, reconocimiento de patrones y clasificación.

De manera formal, si se tiene un conjunto P formado por n puntos

$$P\{p_1, p_2, \cdots p_n\}$$

y un punto de consulta c, debe encontrarse un punto $p \in$ P, tal que

$$D(p,c) \leq D(r,c) \quad \forall r \in P \tag{3.8}$$

donde $D(p,c)$ es la distancia entre el punto p y el punto c.

El método del vecino más cercano es una estrategia de búsqueda muy simple para determinar un sólo patrón de entre los elementos en un conjunto de patrones de referencia, que es el más similar a un patrón de prueba particular. Esta determinación se toma con base en la distancia o medida de similaridad obtenida de la comparación de dos vectores de características. A menor distancia entre patrones, más cercana la vecindad.

El método del vecino más cercano considera al primer vector en el conjunto de patrones de referencia como el vecino más cercano y lo aisla del conjunto tomando en consideración su distancia al vector de prueba. Enseguida considera la distancia entre el vector de prueba y el segundo vector de referencia, si esta distancia es menor que la existente entre el vecino más cercano actual y el vector de prueba entonces el vector de referencia actual reemplaza al vecino más cercano. Este proceso se repite hasta que el último de los vectores de referencia se compara con el vector de prueba en busca de la mínima distancia. El último vecino más cercano es el patrón de referencia cuya distancia al vector de prueba es la más pequeña encontrada. Este es el procedimiento que se aplica cuando las entradas en el diccionario del sistema no están organizadas en modo alguno.

3.6. Los *k*-vecinos más cercanos

El método de k-vecinos es una estrategia de clasificación basada en el método del vecino más cercano. La idea básica consiste en que no solamente el vecino más cercano en un diccionario determine la clase a la cual deba anexarse un vector de características de prueba, sino que haya varios vecinos cercanos al vector de prueba los cuales determinen la

pertenencia del vector de prueba al grupo que tenga mayoría de miembros presentes entre los k vectores más próximos.

De manera formal, si se tiene un conjunto P formado por n puntos

$$P\{p_1, p_2, \cdots p_n\}$$

y un punto de consulta c, debe encontrarse un conjunto de puntos $\{p_a, p_b, \ldots, p_k\} \in P$, tales que sus distancias al punto c son las más pequeñas. En consecuencia $\{p_a, p_b, \ldots, p_k\}$ son los k-vecinos más cercanos a c.

Al emplear este mecanismo de clasificación se debe contar en el diccionario del sistema con varios elementos de la misma clase y todas las clases que el sistema sea capaz de distinguir deben contar con el mismo número de elementos. Para el efecto de la toma de decisión acerca de la clase que tiene mayoría entre los k-vecinos más próximos, es conveniente elegir k como un número impar 3, 5, 7, etc. para que exista posibilidad de romper empate en el número de elementos de cada clase presentes entre los k vecinos próximos. Si k es igual a 1, k-vecinos se reduce al método del vecino más cercano.

Este método opera considerando los primeros k vectores de características en un diccionario como los k vecinos más cercanos, considerando cada una de sus distancias con el vector de prueba. Si la siguiente entrada en el diccionario guarda una distancia menor que alguno de los anteriores k-vecinos, entonces forza la salida del vecino que tenga la mayor distancia al vector de prueba y se inserta en el conjunto. Este procedimiento se repite para todas las entradas en el diccionario, y los vectores que al final permanecen en el conjunto son los k -vecinos que poseen las menores distancias al vector de prueba. Finalmente, se contabilizan los elementos que pertenecen a una misma clase entre los k-vecinos y la clase con mayor cantidad de miembros es a la que se anexa el patrón de prueba.

3.7. Diccionario

Un diccionario es una colección de patrones de referencia que el sistema conoce y los cuales representan señales que están caracterizadas de la misma forma en que se caracteriza una nueva señal de prueba para que ambas puedan ser comparadas. En el caso de los

sistemas de reconocimiento automático de la voz además de los vectores de características se almacena con cada uno de ellos una etiqueta o nombre de clase, que está formada por los símbolos del lenguaje que se emitirán cuando un patrón de prueba se identifique como perteneciente a la clase del patrón o los patrones de referencia que la lleva. En la Tabla 3.1 se proporciona un ejemplo de la organización que pueden tener las entradas en un diccionario para un sistema de reconocimiento automático de voz.

Tabla 3.1: Entrada de diccionario. Palabra *"cero"* caracterizada con sus coeficientes de aproximación de nivel 5 encontrados con la transformada wavelet de Daubechies de orden 2.

Etiqueta	Vector de características
Cero	0.7071 0.7071 0.7068 0.6669 0.8885 0.7808 0.6178 1.0584 0.8567 0.5871 1.0866 0.6973 0.5942 1.1684 0.6684 0.6001 1.2192 0.5463 0.6440 1.2336 0.4417 0.8214 1.1415 0.4829 1.0184 0.8707 0.4261 0.9727 0.8656 0.4874 1.0974 0.7808 0.6023 1.0361 0.8459 0.5922 0.9341 0.9658 0.5027 0.9160 0.9143 0.5210 1.0061 0.6946 0.5655 1.0930 0.6214 0.9059 0.8746 0.7281 0.9045 0.7215 1.0533 0.5311 1.0124 0.6525 0.7872 0.7302 0.9800 0.5622 1.0104 0.6514 0.9112 0.7251 0.7788 0.8144 0.7071 0.7065 0.6965 0.8286 0.7071 0.7071

El uso del criterio del vecino más cercano en la determinación de la mejor coincidencia entre patrones, requiere únicamente de una entrada en el diccionario por cada clase de elocución distinta que se desee identificar, aún cuando se tiene la posibilidad de colocar varias entradas por clase, lo cual incrementa la probabilidad de que el patrón de prueba encuentre uno con suficiente cercanía en la clase correcta y no en otra.

En el caso del criterio de los k-vecinos se requiere que necesariamente ingresen al diccionario varias entradas por cada clase de elocución, de modo tal que exista una alta probabilidad de que la mayoría de las distancias entre los k-patrones más cercanos correspondan a la clase de elocución correcta, y una o varias minorías correspondan a una o varias clases de elocución incorrectas. La decisión toma entonces la forma de una "votación" en la que la mayoría relativa es la triunfadora. Pensemos, por ejemplo, que si de 3-vecinos más cercanos, dos de ellos coinciden en su "opinión", esta sería la que se tomaría en cuenta

en la decisión; es por ello que para este mecanismo no resulta de utilidad el contar con sólo una entrada en el diccionario por clase, dado que todas las posibles opiniones contarían únicamente con el respaldo de una unidad.

El propósito de un sistema automático de reconocimiento de voz queda cubierto al integrar las etapas de medición de distancia entre patrones y clasificación que se han examinado en el presente capítulo, y las etapas precedentes que fueron mencionadas a lo largo del Capítulo 2. Resta entonces la tarea consistente en someter tal sistema a pruebas formales para la evaluación de su comportamiento desde alguna óptica particular, entre las que se pueden mencionar: exactitud, precisión, rapidez, etc. A diferencia del problema de síntesis de voz, en el cual la opinión de las personas (posiblemente expertas en el área: operadores telefónicos y de radio, técnicos controladores de audio, escuchas) pudiese tener un impacto relativamente alto en la determinación de la eficacia del procesamiento aplicado, la evaluación de un sistema de reconocimiento de voz puede efectuarse completamente en forma numérica.

Capítulo 4

Sistema desarrollado

Se construyó un sistema autónomo para el reconocimiento de palabras aisladas utilizando el lenguaje de programación de computadoras en alto nivel Java. Se trata de un lenguaje contemporáneo del que se explotan características avanzadas tales como el procesamiento multihilo y su capacidad para operar en modo de consola o bien sobre vistosas interfaces gráficas de fácil y rápida construcción. Asimismo, se deja abierta la posibilidad de explotar sus ideales prestaciones de integración en red, lo cual puede usarse de ser necesario, por ejemplo, para el envío de comandos verbales a un lugar remoto. Por otra parte, la similitud de su sintaxis con el lenguaje de programación de computadoras de nivel medio C, permite que los algoritmos se traduzcan sin mayor problema a este último si lo que se desea es la construcción, por ejemplo, de un sistema mínimo o bien si lo que se busca es un desempeño que se acerque más al de tiempo real que el que se puede conseguir con un lenguaje interpretado como Java.

En este capítulo se mencionan algunos aspectos de la implantación de los módulos que integran el sistema al cual esencialmente se le puede considerar como un prototipo de laboratorio con el que se lleve a cabo la tarea de experimentación.

El sistema emplea las prestaciones del hardware de una microcomputadora equipada con capacidad multimedia; de esta manera se aprovecha la tarjeta capturadora de audio, a la cual se conecta un micrófono frente al que el usuario pronuncia las palabras.

4.1. Captura de la señal

En lo que respecta a la percepción de la señal acústica generada en la elocución de palabras por una persona, se aprovecha la capacidad multimedia de la mayoría de las computadoras actuales para la adquisición de la señal de voz en formato digital en tiempo real. El estándar de codificación empleado corresponde al PCM (Pulse Code Modulation) sin signo, sobre una tasa de muestreo de 8000 muestras por segundo (8KHz), con una resolución de 8 bits por cada muestra y un solo canal (audio monofónico), características con las que, con una muy alta probabilidad aún el hardware del más bajo rango de prestaciones tendrá la capacidad de abrir una línea de captura de datos.

El valor para la tasa de muestreo de las elocuciones se fija en 8KHz dado que la regla de Nyquist establece que para evitar la confusión de una señal con alguno de sus alias, el muestreo debe realizarse con un ritmo mayor o igual al doble de la frecuencia que posea el componente sinusoidal de frecuencia más alta que integra a la señal. Como en la elocución de palabras no se esperan componentes sinusoidales significativos por encima de los 4KHz, muestrear al doble de esta frecuencia permite efectuar el registro de información que representa con una fidelidad razonable la elocución de una palabra [Lieberman88].

Se encontró que al disponer de un micrófono con alta sensibilidad (no necesariamente un micrófono muy sofisticado, especializado, o de alto costo), con el que bien podrían ya gran cantidad de fabricantes estar equipando computadoras portátiles, aún al deslizar el control de la ganancia de este transductor a un nivel mínimo el dispositivo todavía es capaz de registrar la voz de una persona, evitando el riesgo de saturar la electrónica de conversión analógico/digital y a la vez manteniendo abierta la opción de modificar a la alza la amplitud de la señal de ser esto necesario. Aprovechando esta posibilidad se realizó la normalización del volumen de distintas elocuciones de palabras amplificando señales de magnitud baja, y/o atenuando señales de magnitud alta, ello con el propósito que las diferencias de volumen que se reflejan como amplitud distinta, lo cual para efectos del reconocimiento no es una característica a considerar, no constituyan un factor en la distinción de la información.

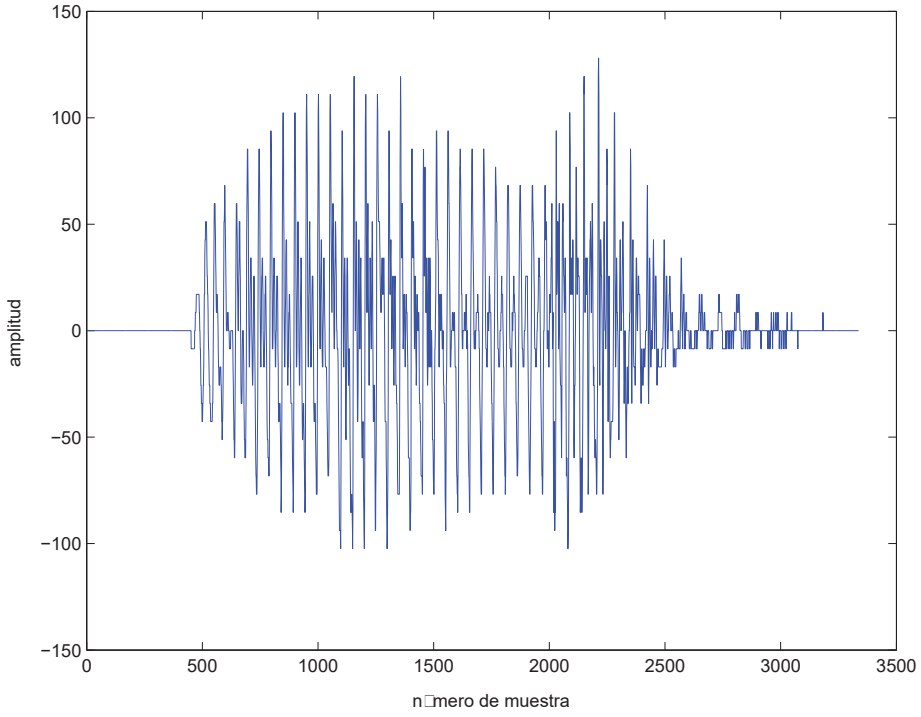

Figura 4.1: Señal obtenida de la etapa de captura del sistema.

4.2. Segmentación de la señal

Si bien es cierto que las perturbaciones acústicas en el ambiente pueden presentar una fuerte dosis de contenidos de alta frecuencia, el desempeño del hardware particular que se utilizó permitió filtrar la mayoría de tales señales sin realizar un procesamiento específico para ello. Un ejemplo del tipo de datos proporcionados por la etapa de captura se muestra en la Figura 4.1, en la que se observa que la señal de entrada permanece con baja magnitud antes de que se inicie la pronunciación de una palabra.

En la Figura 4.2, al hacer un acercamiento en el trazo de la señal sobre la parte del intervalo de tiempo que transcurre desde que se inicia la captura (en la muestra 1) hasta que el usuario comienza a pronunciar (en, aproximadamente, la muestra 320), se revela que el régimen de los cambios de signo por unidad de tiempo es prácticamente nulo o bien despreciable. Estas características que simplifican en gran medida la señal que se estudia, permitieron efectuar la segmentación de la voz implantando un mecanismo sencillo que se

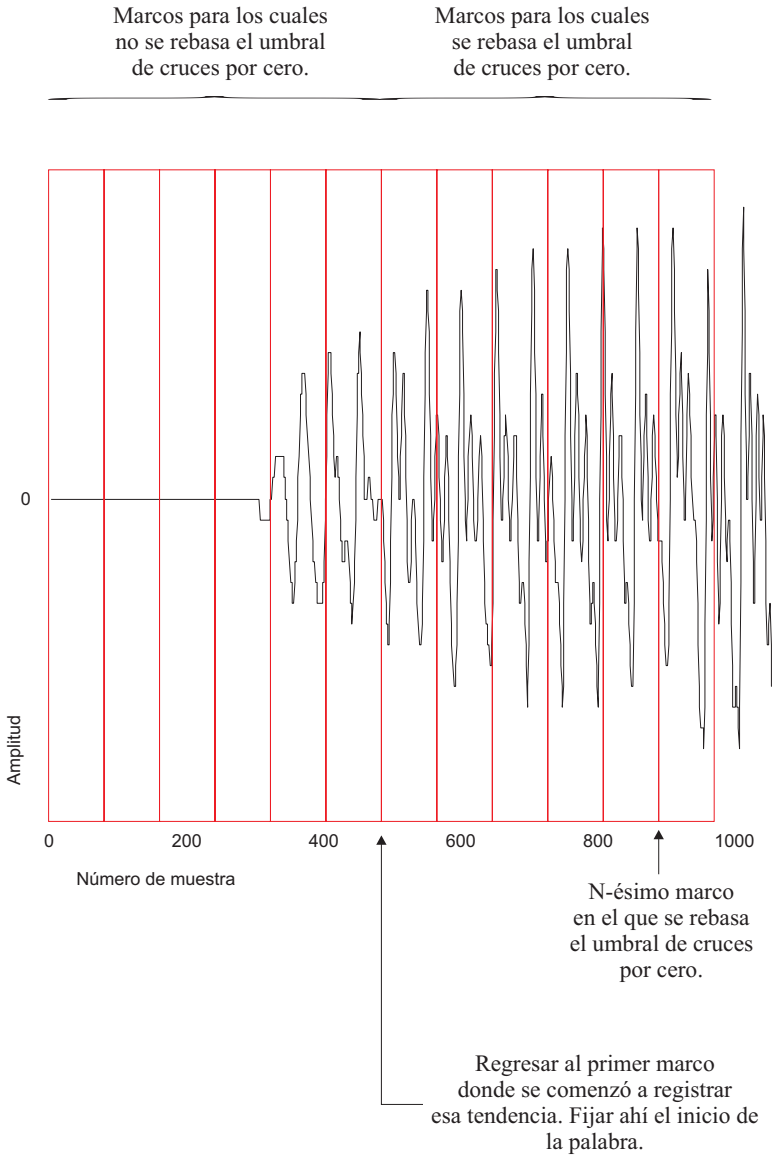

Figura 4.2: Segmentación de la señal con la estrategia del régimen de cruces por cero.

apoya en el régimen de cruces por cero de tiempo corto. La señal se divide en marcos de tiempo correspondientes a 80 muestras, sin traslape entre marcos, y se registra el número de cambios de signo para cada uno de tales intervalos. Si el valor de este número rebasa un cierto umbral determinado experimentalmente durante la fase de sintonización del sistema, y además ese comportamiento se mantiene consistentemente a lo largo de varios marcos consecutivos, se determina entonces el inicio de la palabra a estudiar a partir del marco donde se detectó el principio de esa tendencia. Esto puede entenderse también como el dejar a la señal manifestarse durante cierto tiempo observando la cantidad de cruces por cero que se presentan, y luego "rebobinarla" hasta el punto donde se "recuerda" que comenzaron a aparecer gran cantidad de cruces por cero. Para el caso particular del sistema con el que se experimentó se encuentra que al registrarse dos o más cruces por cero en dos marcos consecutivos de 80 muestras de ancho se tiene un indicio de que con seguridad la palabra puede ya haberse comenzado a pronunciar. La Figura 4.2 ilustra el esquema de análisis de la parte inicial de la señal para la determinación del marco que corresponde al principio de la pronunciación de una palabra.

Como las características al final de la señal capturada son semejantes a las características del inicio de la misma que se han comentado, la determinación del final de la palabra se efectúa con un proceso similar, que comienza la detección de cruces por cero a partir del extremo final del vector de muestras capturadas en dirección hacia el inicio del mismo, es decir, en sentido inverso al utilizado cuando se intentaba determinar el inicio de la palabra.

Dada la existencia de diversas estrategias de segmentación, se cuenta con la posibilidad de utilizar cualquiera de entre ellas por separado (energía de tiempo corto, magnitud promedio de tiempo corto, régimen de cruces por cero de tiempo corto, etc.), o bien la combinación de alguna de las dos primeras de estas con la tercera. Esto es, por ejemplo, se puede sintonizar el sistema para que cuando tanto el número de cruces por cero a lo largo de varios marcos consecutivos analizados, como el nivel de energía de esos mismos marcos rebasan ambos el umbral especificado, se admita que la palabra ha comenzado a pronunciarse a partir del momento en que ambas características se presentaron por primera vez simultáneamente. Este es el esquema que se adoptó en el sistema construido. Por otra

parte, si no se desea rebobinar a un punto previo en la señal se puede optar por el traslape entre marcos analizados, con la finalidad de tener una mejor precisión en la determinación "al vuelo" del inicio o el final de la palabra, es decir, se puede determinar el punto extremo en sitio.

4.3. Extracción de características

Se implantó el algoritmo piramidal de Mallat para calcular la Transformada Discreta con Wavelets, utilizando la función base de Haar y la función base de Daubechies de segundo a décimo orden. La validación del correcto funcionamiento de los algoritmos correspondientes a estos módulos del sistema de reconocimiento se probó exhaustivamente, y se cotejaron los resultados obtenidos de su uso con los que producen los códigos disponibles nativamente en Matlab y las adaptaciones que se codificaron como funciones en Octave a partir del programa propio escrito en Java, encontrándose coincidencia en todos los casos considerados. Se cuenta entonces con un módulo de caracterización de señales confiable que permite construir los vectores de Coeficientes Wavelet de Aproximación en el sistema autónomo que se propone, de manera que no se depende de funciones implantadas en aplicaciones de terceros. Se aprovecha además el sistema en fase de entrenamiento fuera-de-línea, para poder producir las tablas que contienen las entradas del diccionario de palabras que el sistema puede reconocer.

Aún cuando no se pretende que constituya un elemento valioso en la solución del problema del reconocimiento de palabras de acuerdo al enfoque adoptado para el desarrollo de este proyecto, se dotó al sistema de la capacidad de calcular además los Coeficientes de Detalle de la señal, con lo que se implementa la Transformada Wavelet completa.

Los coeficientes de los filtros pasabajas de descomposición de los Wavelets con los cuales de implantó el algoritmo de DWT, se muestran en las Tablas 4.1, 4.2 y 4.3.

Tabla 4.1: Coeficientes de los filtros pasabajas de descomposición de los Waveles de Daubechies de ordenes 1 a 4.

Coeficiente	db01	db02	db03	db04
h0	0.707106781	0.48296291314469	0.33267055295096	0.23037781330886
h1	0.707106781	0.83651630373747	0.80689150931334	0.71484657055254
h2		0.22414386804186	0.45987750211933	0.63088076792959
h3		-0.12940952255092	-0.13501102001039	-0.02798376941698
h4			-0.08544127388224	-0.18703481171888
h5			0.03522629188210	0.03084138183599
h6				0.03288301166698
h7				-0.01059740178500

Tabla 4.2: Coeficientes de los filtros pasabajas de descomposición de los Waveles de Daubechies de ordenes 5 a 7.

Coeficiente	db05	db06	db07
h0	0.16010239797413	0.11154074335008	0.07785205408506
h1	0.60382926979747	0.49462389039839	0.39653931948231
h2	0.72430852843857	0.75113390802158	0.72913209084656
h3	0.13842814590110	0.31525035170924	0.46978228740536
h4	-0.24229488706619	-0.22626469396517	-0.14390600392911
h5	-0.03224486958503	-0.12976686756710	-0.22403618499417
h6	0.07757149384007	0.09750160558708	0.07130921926705
h7	-0.00624149021301	0.02752286553002	0.08061260915107
h8	-0.01258075199902	-0.03158203931803	-0.03802993693503
h9	0.00333572528500	0.00055384220099	-0.01657454163102
h10		0.00477725751101	0.01255099855601
h11		-0.00107730108500	0.00042957797300

Continúa...

Tabla 4.2: (continuación)

Coeficiente	db05	db06	db07
h12			-0.00180164070400
h13			0.00035371380000

Tabla 4.3: Coeficientes de los filtros pasabajas de descomposición de los Waveles de Daubechies de ordenes 8 a 10.

Coeficiente	db08	db09	db10
h0	0.05441584224308	0.03807794736317	0.02667005790095
h1	0.31287159091447	0.24383467463767	0.18817680007762
h2	0.67563073629801	0.60482312367678	0.52720118893092
h3	0.58535468365487	0.65728807803664	0.68845903945259
h4	-0.01582910525602	0.13319738582209	0.28117234366043
h5	-0.28401554296243	-0.29327378327259	-0.24984642432649
h6	0.00047248457400	-0.09684078322088	-0.19594627437660
h7	0.12874742662019	0.14854074933476	0.12736934033574
h8	-0.01736930100202	0.03072568147832	0.09305736460381
h9	-0.04408825393106	-0.06763282905952	-0.07139414716586
h10	0.01398102791702	0.00025094711499	-0.02945753682195
h11	0.00874609404702	0.02236166212352	0.03321267405893
h12	-0.00487035299301	-0.00472320475789	0.00360655356699
h13	-0.00039174037300	-0.00428150368190	-0.01073317548298
h14	0.00067544940600	0.00184764688296	0.00139535174699
h15	-0.00011747678400	0.00023038576400	0.00199240529499
h16		-0.00025196318900	-0.00068585669500
h17		0.00003934732000	-0.00011646685499
h18			0.00009358867000

Continúa. . .

Tabla 4.3: (continuación)

Coeficiente	db08	db09	db10
h19			-0.00001326420300

Los prototipos de los módulos (en Java, los métodos) responsables del cálculo de la DWT son, para la transformada basada en el Wavelet de Haar y los Wavelets de Daubechies respectivamente:

double[][] transformadaWaveletHaar(double[] audioData, char tipo, int nivel)
y
double[][] transformadaWaveletDaubechies(double[] audioData, char tipo, int orden, int nivel)

El primer argumento que se recibe es el vector de datos en la señal cuyo estudio se realiza. El segundo argumento es un carácter ('a' o 'd') que distingue si el análisis requerido es para la obtención de coeficientes de aproximación o de detalle respectivamente. El tercer argumento utilizado exclusivamente en el caso de la transformada con Wavelets de Daubechies indica el número del orden del Wavelet en esta familia. El último argumento en ambos módulos es el número del nivel de profundidad (la escala más alta) que se alcanza en el estudio.

El valor devuelto es una matriz que contiene los coeficientes del tipo solicitado para todos los niveles entre 1 y el que se especificó en el último argumento organizados por renglones, haciendo corresponder el número de renglón con los distintos niveles en que se efectuó la transformación, esto es: el renglón 1 contiene los coeficientes de nivel 1, el renglón 2 contiene los coeficientes de nivel 2 y así sucesivamente.

4.4. Medición de distancia

Se implantó el algoritmo de Doblado Dinámico en Tiempo para alineación temporal de patrones de distinta longitud, con la expectativa de que la distancia global resultante

del trayecto de alineación brinde una medida de similaridad adecuada. Un mecanismo simple que permite efectuar la medición de la separación entre dos vectores se muestra en el Algoritmo 3.

Algoritmo 3 Doblado Dinámico en Tiempo

```
DOBLADODINAMICOENTIEMPO(vector a, vector b)
 1  n ← longitudDe : a
 2  m ← longitudDe : b
 3  D[1][1] ← distancia(1, 1)
 4  para j = 2 hasta m
 5    D[1][j] ← D[1][j − 1] + distancia(a[1], b[j])
 6    para i = 2 hasta n
 7    D[i][1] ← D[i − 1][1] + distancia(a[i], b[1])
 8    k ← 2
 9  mientras k <= n, k <= m
10    para i = k hasta n
11      d ← distancia(a[i], b[k])
12      D[i][k] ← minimo(D[i − 1][k − 1] + d, D[i − 1][k] + 2 ∗ d, D[i][k − 1] + 2 ∗ d)
13      para j = k + 1 hasta m
14      d ← distancia(a[k], r[j])
15      D[k][j] ← minimo(D[k − 1][j − 1] + d, D[k − 1][j] + 2 ∗ d, D[k][j − 1] + 2 ∗ d)
16      regresaD[n][m]/(n + m)
```

4.5. Clasificación

En lo que respecta al vocabulario del sistema, se construyeron diccionarios para Transformaciones Wavelet aplicadas en 5 diferentes escalas conservando únicamente los

coeficientes de aproximación. El vector de características correspondiente a cada uno de los ejemplares que ingresa a los distintos diccionarios, ejemplo del cual se presentó en la Sección 3.7, tiene longitudes típicas para palabras cortas (se habla de palabras tales como los dígitos con pronunciaciones de alrededor de unos 300ms de duración) que varían de acuerdo con la escala de transformación de la cual ha sido producto. Estas magnitudes van desde aproximadamente 1200 elementos para caracterización en escala 1, hasta alrededor de 80 elementos cuando la caracterización se efectúa en escala 5, independientemente del tipo de Wavelet utilizado en la transformación. Las pruebas que se mencionan en el capítulo siguiente enfatizan en el aprovechamiento de vectores de características con alrededor de 80, 160 o 320 elementos en el peor de los casos para cada ejemplar. Las comparaciones se efectúan de manera homogénea, es decir, al cotejar un patrón de prueba con uno de referencia ello se realiza considerando la misma escala para ambos.

Se cuenta pues con un sistema sencillo pero completo y autónomo para el reconocimiento de palabras, en el que se ha enfatizado el desarrollo del módulo de caracterización de señales (o de extracción de características) al aplicar una técnica de filtrado distinta a las clásicas, lo cual se ha integrado con los módulos restantes del sistema en los que se emplean técnicas ampliamente difundidas y extensamente probadas con éxito. Una de las virtudes del sistema con el que se cuenta, es que simplemente con conocer los coeficientes del filtro pasabajas para otro Wavelet además de los aquí considerados, y haciendo el intercambio con los existentes en el módulo extractor de características técnicamente se cuenta con un sistema nuevo y distinto de los anteriores, por las propiedades del Wavelet en el que se basa, y así las posibilidades de experimentación se amplían considerablemente. En el capítulo siguiente se muestran las indicaciones de cómo diseñar algunos ensayos de reconocimiento interesantes, lo cual puede marcar la pauta para detectar Wavelets con mejores atributos desde el punto de vista del reconocimiento de voz. Asimismo, se emplea una metodología de evaluación del rendimiento de un sistema de este tipo, que se considera es adecuada dado que se basa en la interpretación geométrica de métricas, con lo que las tendencias de funcionamiento son fácilmente detectables.

Capítulo 5

Resultados

5.1. Gráficas ROC (Receiver Operation Characteristic)

Las gráficas *Característica de Operación de Receptor* (ROC por sus siglas en inglés) son una técnica útil para la visualización, la organización y la selección de clasificadores basada en su rendimiento. Las gráficas ROC se usan comúnmente en la toma de decisiones médicas y se han adoptado de manera creciente en las comunidades que investigan el aprendizaje en máquinas y la minería de datos. Esta técnica se ha utilizado por largo tiempo en el área de *Teoría de Detección de Señales* (SDT, Signal Detection Theory) para revelar el equilibrio entre las tasas de acierto y las tasas de falsa alarma de un clasificador. [Spackman89] al adoptar las gráficas ROC en el campo del aprendizaje en máquinas demostró el valor que tienen estas en la evaluación y la comparación de algoritmos.

5.1.1. Rendimiento de un clasificador

Algunos problemas de clasificación emplean únicamente dos clases. En lo formal, cada instancia I miembro de una clase se mapea a un elemento del conjunto $\{p, n\}$ de etiquetas de clase positiva y negativa respectivamente. Un *modelo de clasificación* (o *clasificador*) es un mapeo desde las instancias hacia las clases predichas. Algunos modelos de clasificación producen una salida continua (es decir, una estimación de probabilidad de pertenencia a una clase para una instancia). Otros modelos producen una etiqueta discreta de clase la

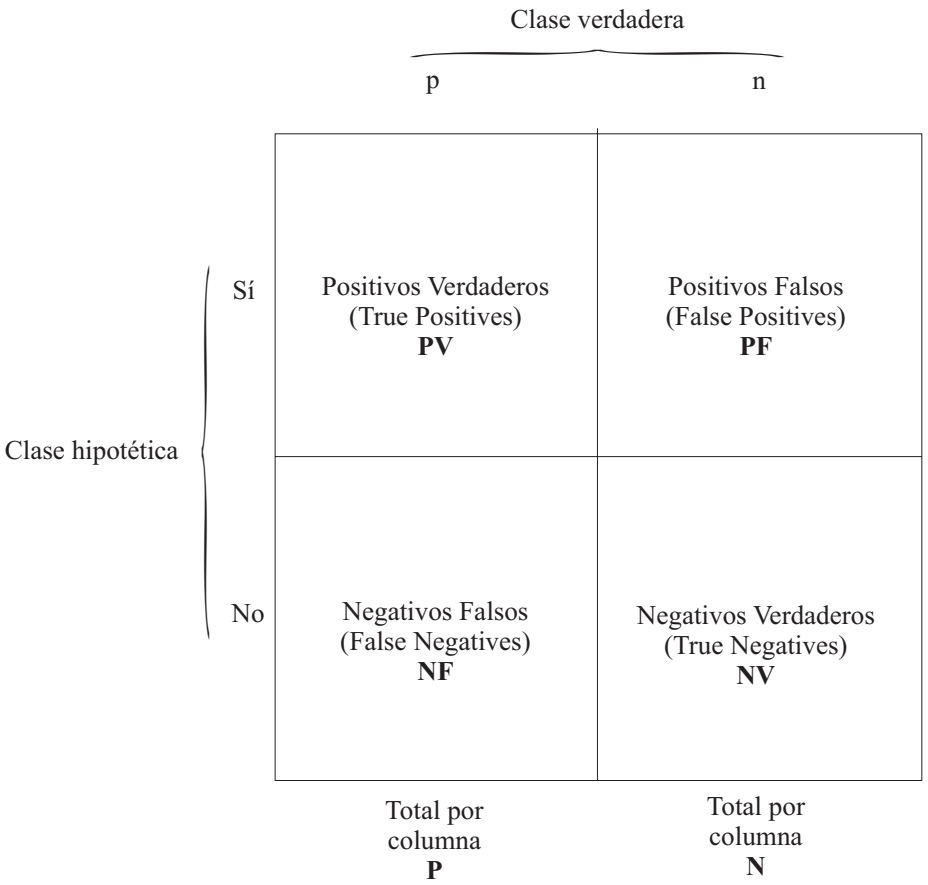

Figura 5.1: Matriz de confusión.

cual indica solamente la clase predicha de la instancia. Para efectuar la distinción entre la clase real y la clase predicha, se emplean las etiquetas $\{Sí, No\}$ para las predicciones de clase que el modelo produce. Dado un clasificador y una instancia, existen cuatro posibles resultados. Si la instancia es positiva y es clasificada como positiva se le cuenta como un *positivo verdadero, (PV)*; si es clasificada como negativa se le cuenta como un *negativo falso, (NF)*. Si la instancia es negativa y es clasificada como negativa, se le cuenta como un *negativo verdadero, (NV)*; si es clasificada como positiva, se le cuenta como un *positivo falso, (PF)*. Dado un clasificador y un conjunto de instancias (el conjunto de prueba), se puede construir una *matriz de confusión* de dos-por-dos (también llamada tabla de contingencia), la cual representa las disposiciones del conjunto de instancias. Esta matriz forma la base de muchas métricas comunes.

En la Figura 5.1 se muestra una matriz de confusión a partir de la cual se derivan las ecuaciones de varias métricas comunes que pueden calcularse. Los números a lo largo de la diagonal principal representan las decisiones tomadas correctamente, y los números fuera de la diagonal representan los errores (la confusión) entre las varias clases. Conceptualmente, la tasa de positivos verdaderos (también llamada *tasa de aciertos*, asimismo *recuperación*, o en inglés *true positive rate*, *hit rate*, *recall* respectivamente) de un clasificador se estima de acuerdo con 5.1.

$$\textit{tasa de positivos verdaderos} \approx \frac{\text{Positivos clasificados correctamente}}{\text{Positivos en total}} \tag{5.1}$$

Conceptualmente, la tasa de positivos falsos (también denominada *tasa de falsas alarmas*, o en inglés *false positive rate*) del clasificador se calcula de acuerdo con 5.2.

$$\textit{tasa de positivos falsos} \approx \frac{\text{Negativos clasificados incorrectamente}}{\text{Negativos en total}} \tag{5.2}$$

En la Figura 5.1 debe tenerse cuidado de no confundir las etiquetas $\{p, n\}$ positiva y negativa respectivamente de instancias de la clase real que se escriben usando letras minúsculas y aparecen en la parte superior de la matriz, con las cantidades P y N que aparecen en la parte inferior de la matriz usando letras mayúsculas y que representan los totales calculados por columna, es decir:

$$P = PV + \text{NF} \tag{5.3}$$

y

$$N = PF + \text{NV} \tag{5.4}$$

Algunos otros términos asociados con las curvas ROC son:

$$\textit{sensitividad} = \text{recuperación} = \textit{tasa de positivos verdaderos} \tag{5.5}$$

$$especificidad = \frac{\text{Negativos Verdaderos}}{\text{Positivos Falsos} + \text{Negativos Verdaderos}} = 1 - tasa\ de\ positivos\ falsos \tag{5.6}$$

$$valor\ predictivo\ positivo = \text{precisión} \tag{5.7}$$

De acuerdo con los elementos en la matriz de confusión:

$$tasa\ de\ positivos\ falsos = \frac{PF}{N} \tag{5.8}$$

$$tasa\ de\ positivos\ verdaderos = \frac{PV}{P} \tag{5.9}$$

$$precisión = \frac{PV}{PV + PF} \tag{5.10}$$

$$recuperación = \frac{PV}{P} \tag{5.11}$$

$$precisión = \frac{PV + NV}{P + N} \tag{5.12}$$

$$medición\text{-}F = \frac{2}{1/precisión + 1/recuperación} \tag{5.13}$$

5.1.2. Espacio ROC

Las gráficas ROC son gráficos bidimensionales en los cuales la tasa de positivos verdaderos se dibuja en el eje Y y la tasa de positivos falsos se dibuja en el eje X. Un gráfico ROC revela el equilibrio entre los beneficios (positivos verdaderos) y los costos (los positivos falsos). La Figura 5.2 muestra una gráfica ROC con cinco clasificadores identificados como A hasta E.

Un clasificador *discreto* es aquel que emite solo una etiqueta de clase. Cada clasificador discreto produce un par (*tasa de positivos falsos*, *tasa de positivos verdaderos*)

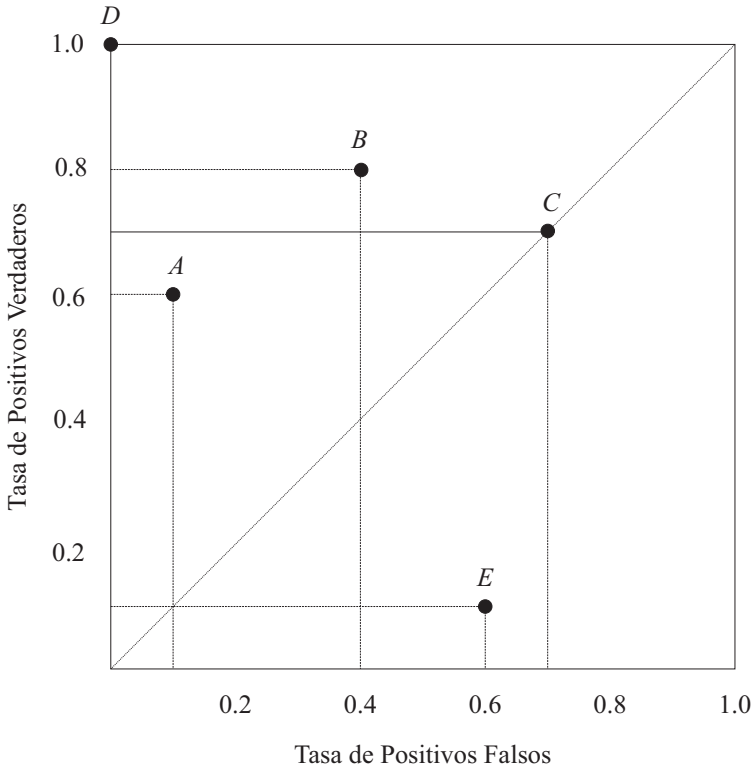

Figura 5.2: Distintos clasificadores en el espacio ROC.

correspondiente a un único punto en el espacio ROC. Todos los clasificadores en la Figura 5.2 son clasificadores discretos. En el espacio ROC se destacan varios puntos importantes. La esquina inferior izquierda (0,0) representa la estrategia de no emitir nunca una clasificación positiva; un clasificador como tal no comete errores por positivos falsos pero tampoco gana en positivos verdaderos. La estrategia opuesta, de emitir clasificaciones positivas incondicionalmente se representa por el punto en la esquina superior derecha (1,1).

El punto (0,1) representa la clasificación perfecta. El rendimiento del clasificador D es perfecto, como la gráfica lo muestra. Informalmente, un punto dentro del espacio ROC es mejor que otro si se ubica al noroeste (la tasa de positivos verdaderos es mayor, la tasa de positivos falsos es menor, o ambas cosas) del primero. Los clasificadores que aparecen en el lado izquierdo de una gráfica ROC, cerca de la línea del eje Y, pueden considerarse como "conservadores"; estos efectúan clasificaciones positivas solamente bajo fuerte evidencia de manera que cometen pocos errores de positivo falso, pero a menudo también tienen

bajas tasas de positivos verdaderos. Los clasificadores en el lado superior derecho de una gráfica ROC pueden considerarse como "liberales"; estos realizan clasificaciones positivas con evidencia débil de tal forma que clasifican casi todos los positivos correctamente, pero a menudo tienen también altas tasas de positivos falsos. En la Figura 5.2 el clasificador A es más conservador que el clasificador B. Muchos dominios del mundo real se rigen predominantemente por instancias negativas, y de este modo, el rendimiento en el extremo izquierdo del gráfico ROC se torna más interesante.

La línea diagonal $y = x$ representa la estrategia de suponer aleatoriamente una clase. Por ejemplo, si un clasificador supone aleatoriamente la clase positiva la mitad del tiempo, puede esperarse que se obtengan la mitad de los positivos y la mitad de los negativos correctos; esto produce el punto (0.5,0.5) en el espacio ROC. Si el clasificador supone la clase positiva 90 % de las ocasiones, se puede esperar que obtenga 90 % de los positivos correctamente, pero su tasa de positivos falsos se incrementará a 90 % también, produciendo el punto (0.9,0.9) en el espacio ROC. De este modo, un clasificador aleatorio producirá un punto ROC que se "desliza" adelante y atrás sobre la diagonal, basándose en la frecuencia con la cual supone la clase positiva. Con el objeto de alejarse de esta diagonal e internarse en la región triangular superior, el clasificador deberá explotar alguna información contenida en los datos. En la Figura 5.2 el rendimiento del clasificador C es virtualmente aleatorio. En (0.7,0.7) se puede decir que C está suponiendo la clase positiva 70 % del tiempo. Cualquier clasificador que aparece en el triángulo inferior tiene un rendimiento peor que lo que resulta del hecho de suponer la clase positiva aleatoriamente. De este modo, ese triángulo siempre debe estar vacío en las gráficas ROC. Se puede decir de cualquier clasificador en la diagonal, que no tiene información acerca de la clase. Respecto de un clasificador abajo de la diagonal tal como es el caso del clasificador E, se puede decir que dispone de información útil, pero que la está aplicando o interpretando incorrectamente, y por lo tanto, nunca debería aparecer un clasificador ubicado en el triángulo inferior del gráfico. De hecho, el clasificador E se puede tornar en el clasificador A, simplemente con el uso a la inversa de la evidencia de la que dispone.

Muchos clasificadores están diseñados para producir solamente una decisión de clase, es decir, para emitir un *Sí* o *No* sobre cada instancia. Cuando se le aplica un clasificador

como tal a un conjunto de prueba, ello produce una única matriz de confusión, lo que a su vez corresponde a un único punto ROC. De este modo, un clasificador discreto produce un solo punto en el espacio ROC. Para generar un conjunto de puntos que permitan el trazado de una curva, puede utilizarse un umbral que determine el resultado de la clasificación: si la salida del clasificador está por encima del umbral, el clasificador produce un *Sí*, de otro modo produce un *No*. Cada valor de umbral produce un punto en el espacio ROC. Conceptualmente, podemos imaginar al umbral variando desde $-\infty$ hasta $+\infty$ y efectuando con ello el trazado de una curva dentro del espacio ROC [Fawcett04].

5.2. Experimentos realizados

Se realizaron diversos experimentos consistentes en la elocución de las palabras "uno", "dos", "tres", "cuatro", "cinco", "seis", "siete", "ocho", "nueve", "cero" en diez instancias cada una. La caracterización de la señal se lleva a cabo mediante un vector unidimensional que contiene los coeficientes de aproximación producto de la transformada wavelet en cada uno de los primeros cinco niveles de descomposición (en las escalas 1 a 5). Posteriormente se examinó la distancia existente entre la primera elocución de la palabra "uno" consigo misma (distancia igual a cero), y con las 99 elocuciones del resto de las palabras en un mismo nivel de descomposición. Una vez hecho esto, se procedió al ordenamiento creciente de los valores de las distancias calculadas. Al repetir este procedimiento para distintos tipos de wavelet se puede encontrar el comportamiento que demuestran las distintas opciones de caracterización correspondientes a los distintos tipos de wavelets empleados en la etapa correspondiente del proceso de reconocimiento, al efectuarse la determinación del número necesario de vecinos más cercanos necesarios de ser considerados para lograr una identificación positiva de una palabra. Se condujeron experimentos similares al que se describió anteriormente, en el que los vectores de características contienen a los coeficientes de detalle producto de la transformada wavelet en cada uno de los primeros cinco niveles de descomposición.

5.3. Caracterización de palabras mediante coeficientes de aproximación obtenidos con transformada basada en el wavelet de Haar (db01).

Las distancias encontradas entre elocuciones de 100 palabras correspondientes a los dígitos se pueden consultar en la Figura 5.3.

A primera vista, de entre todas las distancias entre las palabras se observan valores más pequeños para la asociación de la palabra pronunciada ("uno") y varias de las que pertenecen a su misma clase (otros "uno"s), esto es, palabras que corresponden a una identificación positiva en el reconocimiento. Para contar con una evidencia más en este mismo sentido, se efectúa el cálculo del promedio de las distancias entre la palabra pronunciada y todas las que pertenecen a una misma clase para todas y cada una de las clases conocidas (en este caso, una clase para cada dígito), evitando incluir en el cálculo del primero de los promedios (para la clase de la palabra "uno") la distancia nula, que corresponde una misma instancia de la palabra "uno", esto es porque no se comparan dos elocuciones distintas. Así, el promedio de distancia para la palabra "uno" se obtiene considerando 9 muestras y los promedios para las palabras restantes correspondientes a los dígitos se obtienen considerando 10 muestras. Los resultados de esta prueba se pueden consultar en la Figura 5.4.

Se percibe que el promedio de distancias más pequeño, es decir, la más alta cercanía entre palabras, se presenta entre la palabra pronunciada y las otras palabras en su propia clase, mientras que las distancias promedio al resto de las clases a las cuales la palabra pronunciada no pertenece, son de magnitudes mayores.

En la Tabla 5.1 se muestra el número de los primeros 28 vecinos más cercanos a la palabra "uno" del total de 99 patrones de referencia, de acuerdo con su distancia. Se indica además entre paréntesis la clase a la que cada uno de ellos corresponde. La fila sombreada corresponde a la aparición del primer vecino que corresponde a una clase distinta a la de la palabra "uno". Del análisis de estos resultados se encuentra que los primeros 6 vecinos más cercanos del total de 99 hacen corresponder la elocución de la palabra "uno" con palabras en su clase (otras palabras "uno"), y que no es sino hasta considerados los primeros 28 vecinos más cercanos del total de 99, lo que implica la aparición de una asociación de la

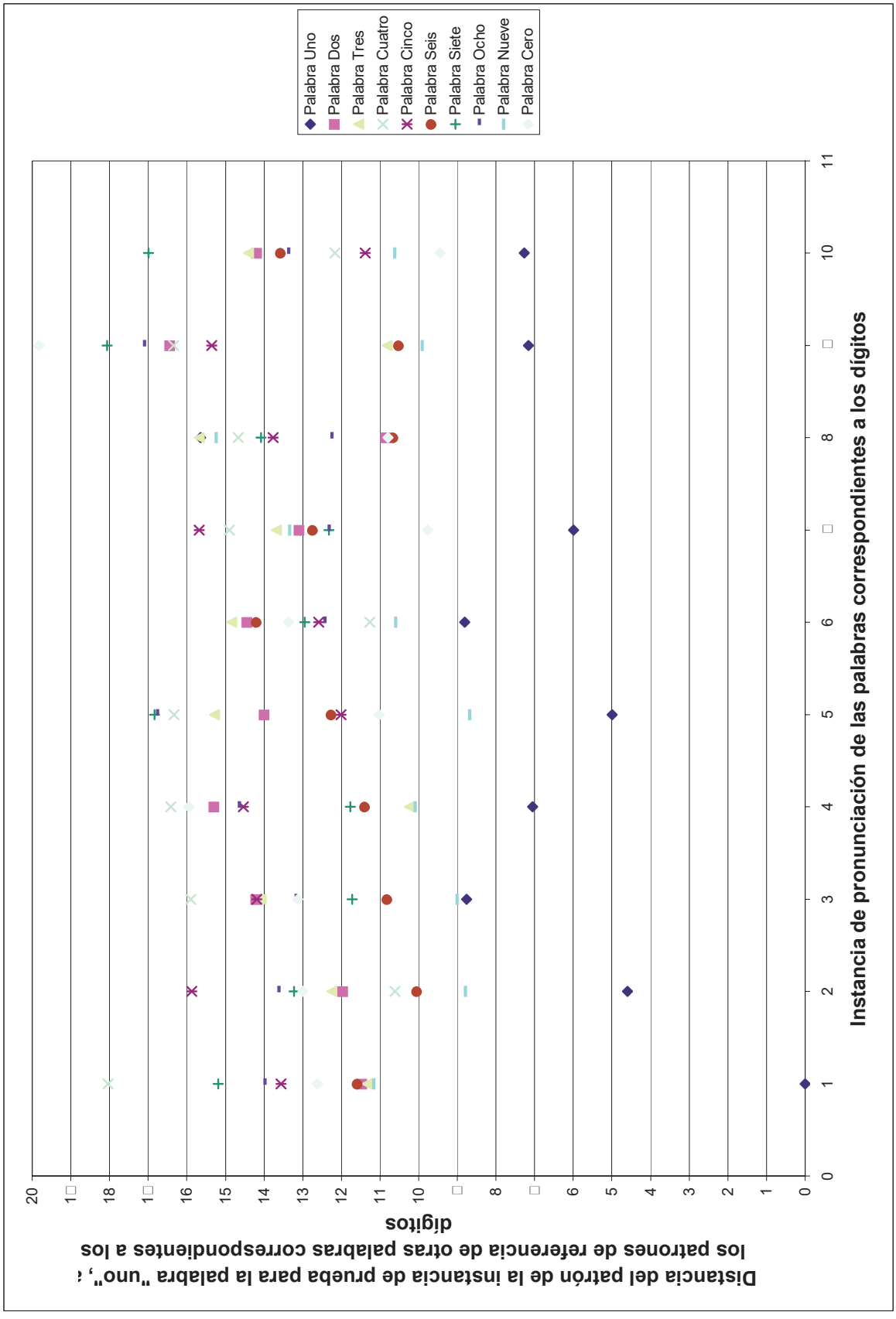

Figura 5.3: Distancias entre los patrones de 10 instancias de pronunciación de 10 clases distintas palabras y el patrón de prueba correspondiente a la elocución de la palabra *"uno"*.

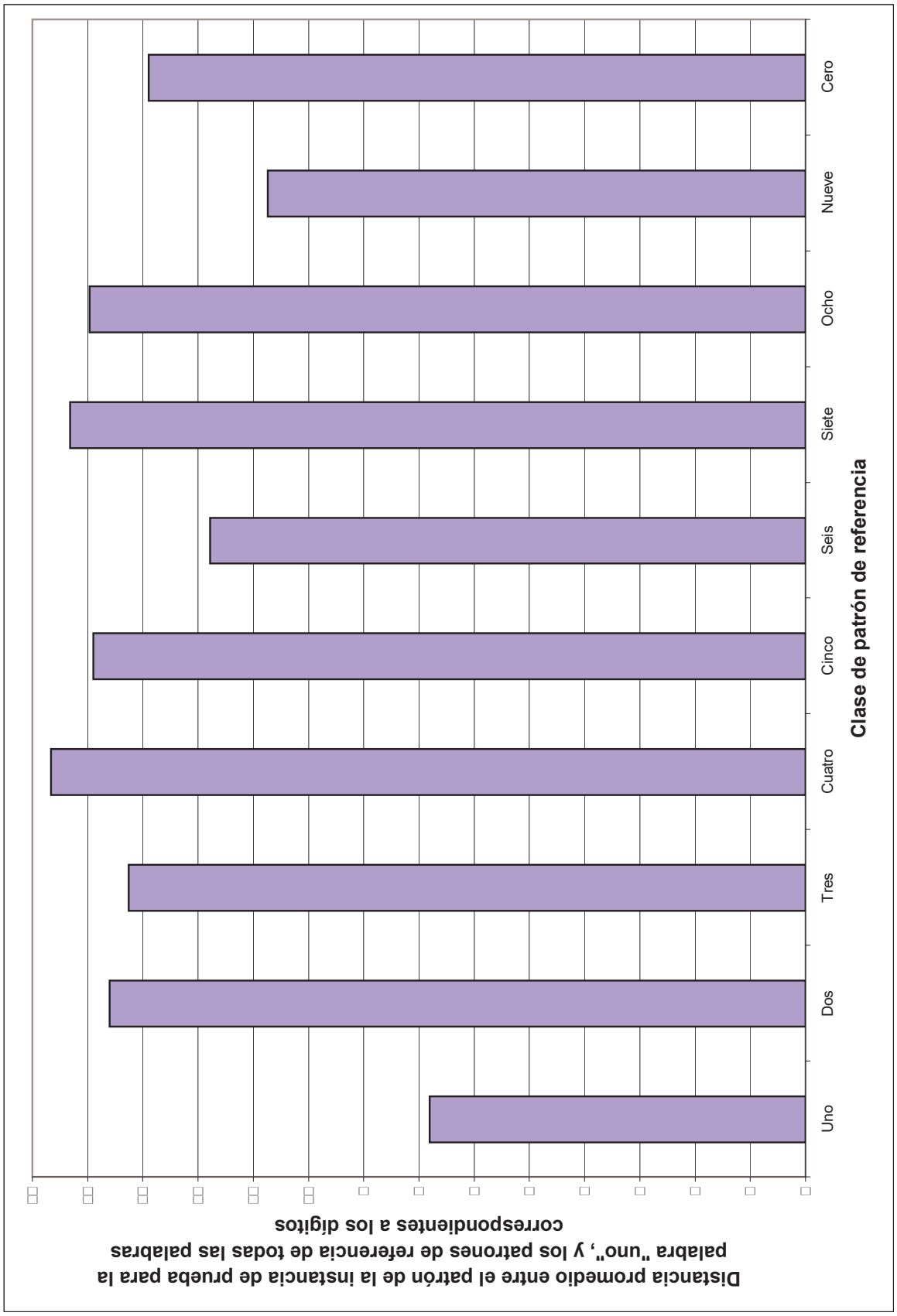

Figura 5.4: Distancias promedio entre los patrones de 10 instancias de pronunciación de 10 clases distintas palabras y el patrón de prueba correspondiente a la elocución de la palabra *"uno"*.

palabra con vecinos pertenecientes a otras 6 clases de palabras incorrectas de un total de 10 clases de distintas palabras consideradas, que se produce una paridad de "opiniones" entre dos clases diferentes.

Tabla 5.1: Primeros 28 vecinos más cercanos en la elocución de la palabra *"uno"* y 99 otras elocuciones de todas las palabras correspondientes a los dígitos, *"uno"* inclusive, caracterizadas con wavelet de Haar en el primer nivel de descomposición.

		No pertenecientes a una					
Número de vecinos	Pertenecientes a la clase correcta	1a clase (9)	2a clase (0)	3a clase (6)	4a clase (3)	5a clase (4)	6a clase (2)
1	1	0	0	0	0	0	0
2	2	0	0	0	0	0	0
3	3	0	0	0	0	0	0
4	4	0	0	0	0	0	0
5	5	0	0	0	0	0	0
6	6	0	0	0	0	0	0
7	6	1	0	0	0	0	0
8	7	1	0	0	0	0	0
9	7	2	0	0	0	0	0
10	8	2	0	0	0	0	0
11	8	3	0	0	0	0	0
12	8	3	1	0	0	0	0
13	8	3	2	0	0	0	0
14	8	4	2	0	0	0	0
15	8	4	2	1	0	0	0
16	8	5	2	1	0	0	0

Continúa...

Tabla 5.1: (Continuación)

		No pertenecientes a una					
Número de vecinos	Pertenecientes a la clase correcta	1a clase (9)	2a clase (0)	3a clase (6)	4a clase (3)	5a clase (4)	6a clase (2)
17	8	5	2	1	1	0	0
18	8	5	2	2	1	0	0
19	8	6	2	2	1	0	0
20	8	6	2	2	1	1	0
21	8	7	2	2	1	1	0
22	8	7	2	3	1	1	0
23	8	7	3	3	1	1	0
24	8	7	3	3	2	1	0
25	8	7	3	4	2	1	0
26	8	7	3	4	2	1	1
27	8	7	4	4	2	1	1
28	8	8	4	4	2	1	1

La Figura 5.5 presenta las curvas ROC correspondientes al funcionamiento de cinco clasificadores basados en la caracterización de los datos mediante Wavelet de Haar en cinco distintas escalas de descomposición, considerando los patrones correspondientes a 10 elocuciones de cada uno de los diez dígitos. Destaca a primera vista la cercanía de las líneas con la esquina superior izquierda del gráfico, lo cual denota altas tasas de reconocimiento.

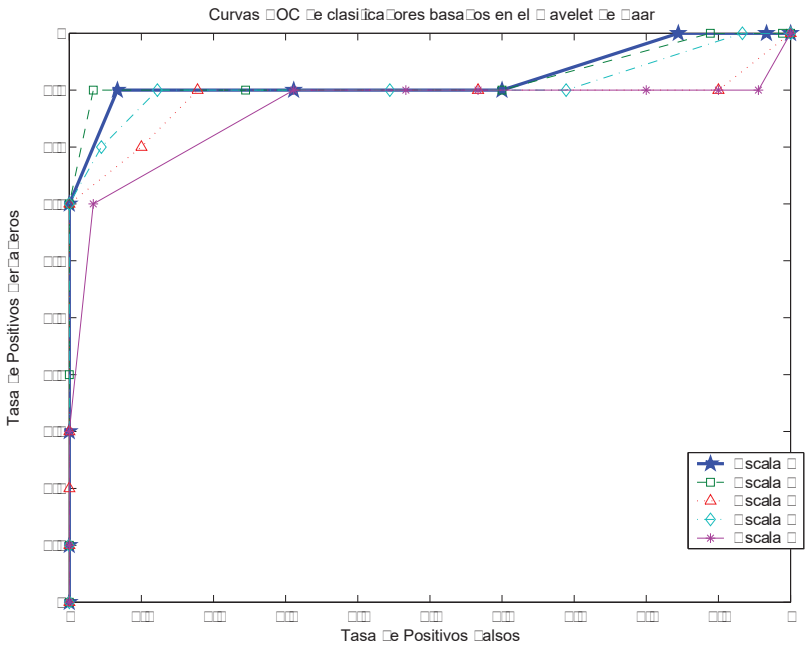

Figura 5.5: Curvas ROC para cinco clasificadores basados en el Wavelet de Haar.

5.4. Caracterización de palabras mediante coeficientes de aproximación obtenidos con transformada basada en el wavelet de Daubechies (db02 a db10).

Las curvas ROC que se generan para los clasificadores basados en el wavelet de Daubechies de 2^o a 10^o ordenes en las primeras cinco escalas de descomposición, considerando los patrones correspondientes a 10 elocuciones de cada uno de los diez dígitos se muestran en las Figuras 5.6 a 5.14 respectivamente.

5.5. Características de los clasificadores a partir de la información revelada por las curvas ROC.

El primer aspecto destacado en el comportamiento de las curvas obtenidas es que la superficie que delimitan es mayor a 0.5 en la totalidad de ellas. Esto habla acerca de que la sensibilidad de los clasificadores, es decir, su aptitud para discriminar entre la clase correcta

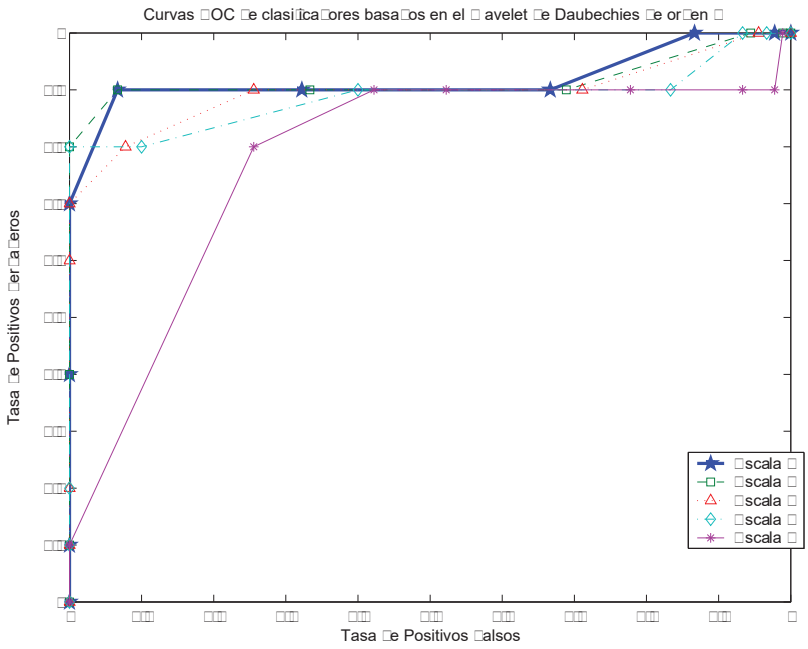

Figura 5.6: Curvas ROC para cinco clasificadores basados en el Wavelet de Daubechies de orden 2.

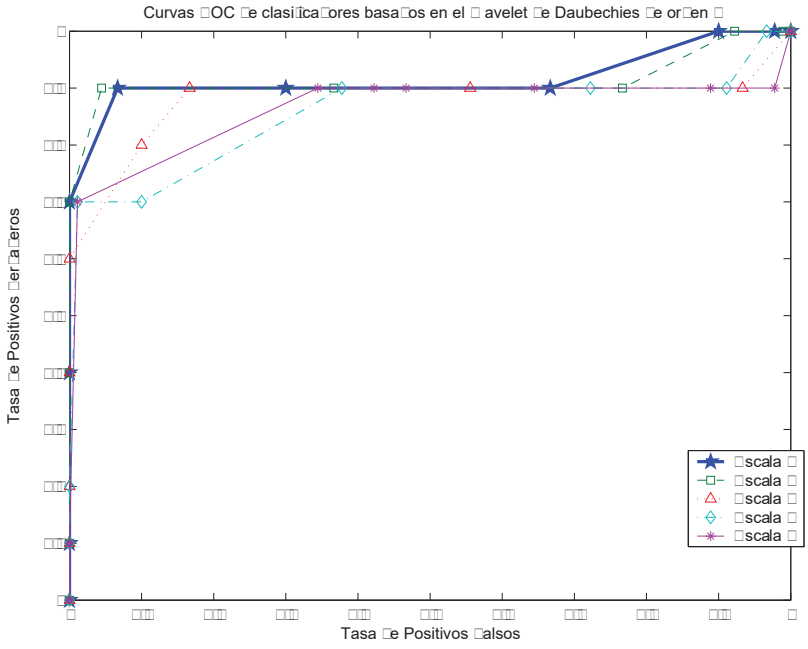

Figura 5.7: Curvas ROC para cinco clasificadores basados en el Wavelet de Daubechies de orden 3.

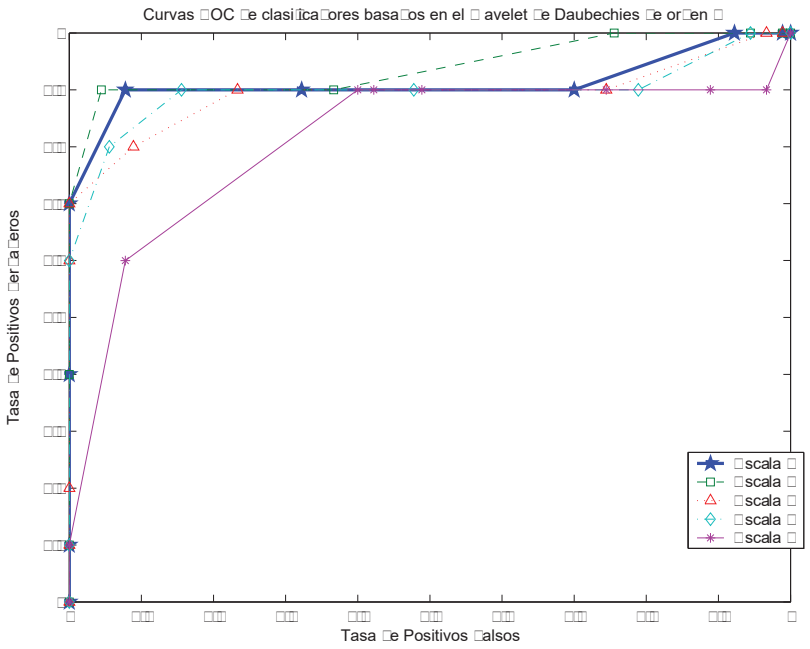

Figura 5.8: Curvas ROC para cinco clasificadores basados en el Wavelet de Daubechies de orden 4.

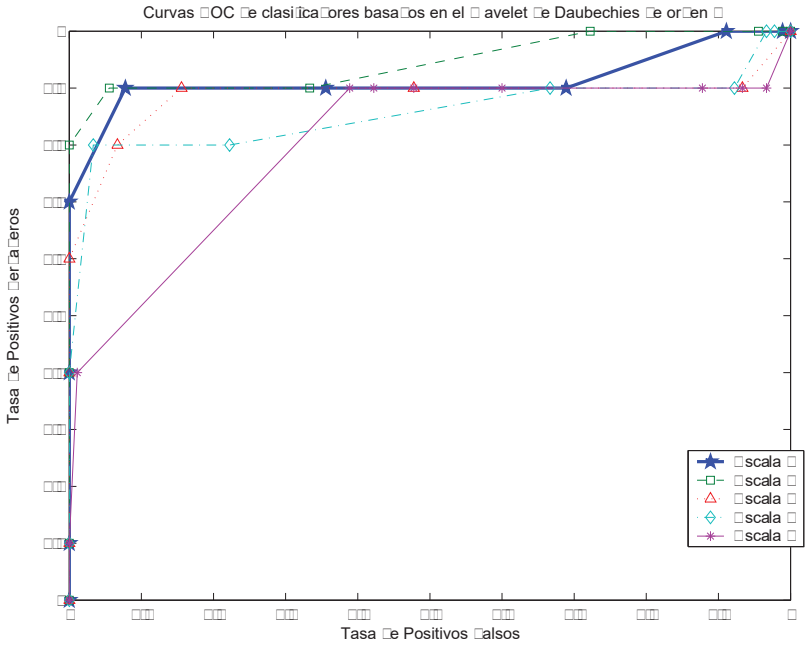

Figura 5.9: Curvas ROC para cinco clasificadores basados en el Wavelet de Daubechies de orden 5.

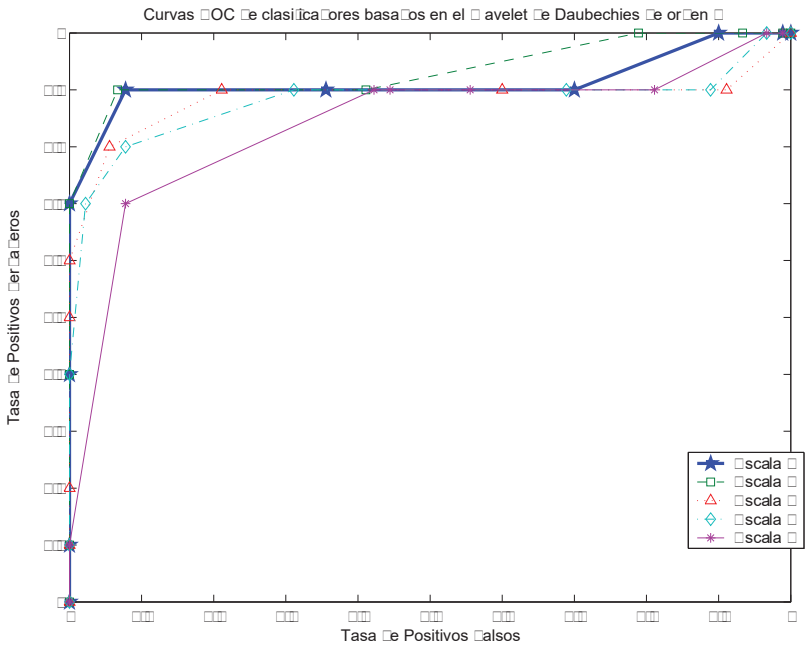

Figura 5.10: Curvas ROC para cinco clasificadores basados en el Wavelet de Daubechies de orden 6.

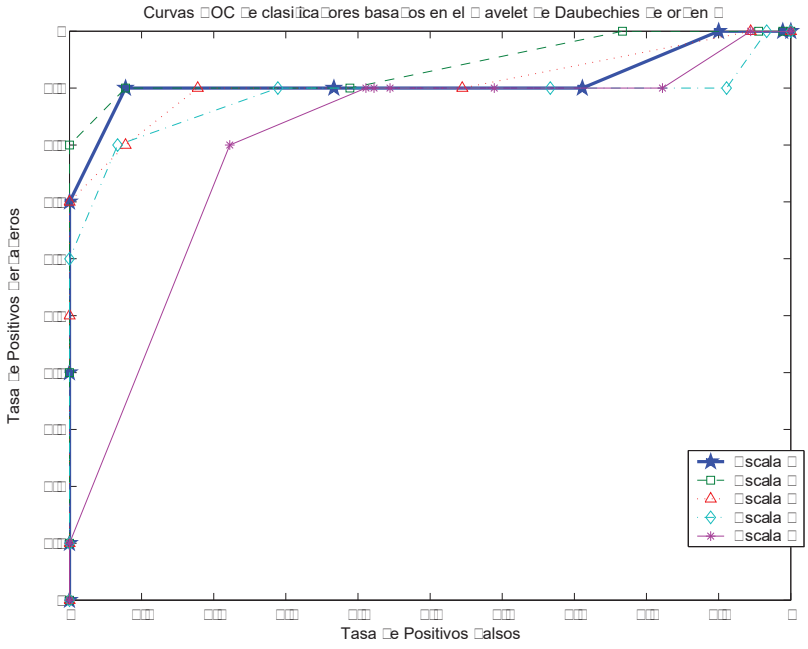

Figura 5.11: Curvas ROC para cinco clasificadores basados en el Wavelet de Daubechies de orden 7.

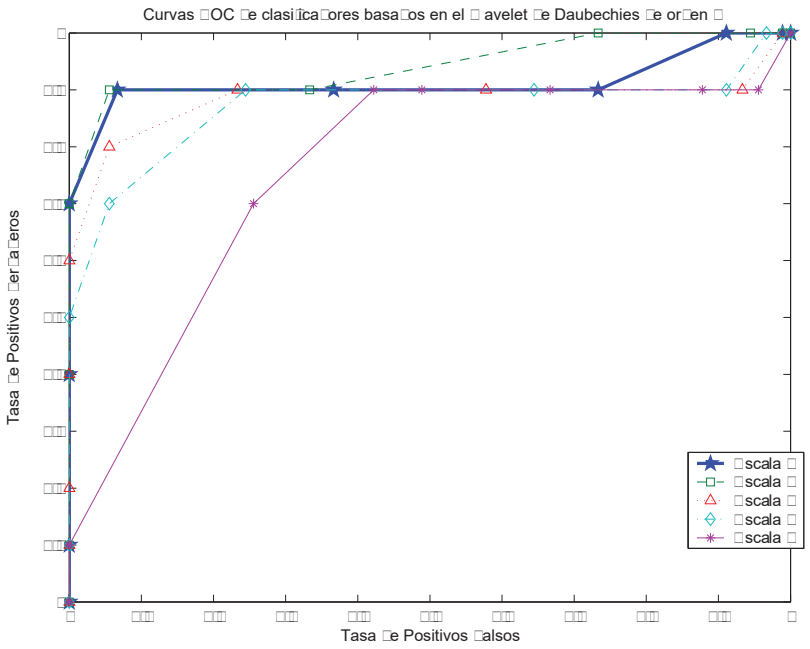

Figura 5.12: Curvas ROC para cinco clasificadores basados en el Wavelet de Daubechies de orden 8.

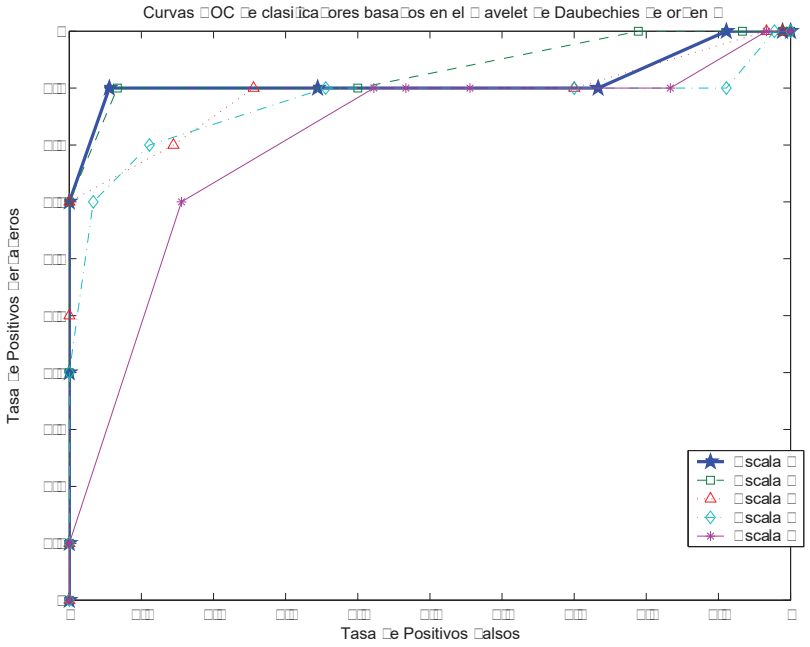

Figura 5.13: Curvas ROC para cinco clasificadores basados en el Wavelet de Daubechies de orden 9.

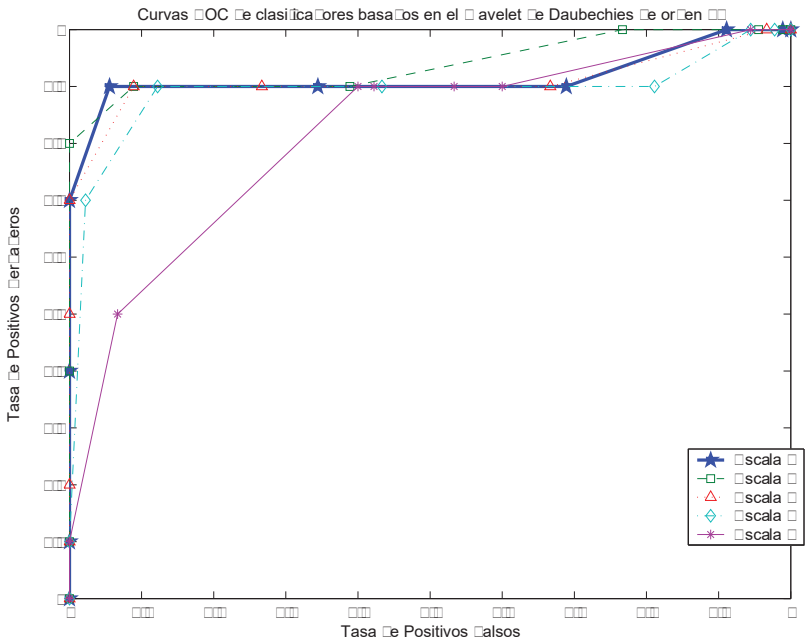

Figura 5.14: Curvas ROC para cinco clasificadores basados en el Wavelet de Daubechies de orden 10.

y el resto de las clases incorrectas es alta. Se entiende entonces que los clasificadores han detectado en la estructura de los datos información pertinente y útil para el reconocimiento.

De acuerdo con lo esperado, los clasificadores que más se acercan a la esquina superior izquierda en los espacios ROC de las Figuras 5.5 a 5.14 son aquellos basados en extracción de características con Wavelets en transformaciones de escalas bajas: 1 y 2, para las que aún persisten cantidades significativas de rasgos de alta frecuencia junto a los rasgos de baja frecuencia en el resultado de la caracterización de la señal de voz. Esto resultó de la misma forma para todo clasificador basado en transformaciones Wavelet en la primeras dos escalas, independientemente del tipo de Wavelet empleado.

No obstante que este último hallazgo sugiera la posibilidad de construir un sistema de reconocimiento con altas tasas de acierto basado en ese enfoque particular, debe recordarse que la baja escala de los Wavelets usados en la transformación trae consigo la necesidad de manipular volúmenes altos de datos, ya que cada elemento en la caracterización de escala 1 representa solamente a cada dos datos en la señal original, en tanto que para la caracterización de escala 2 cada elemento en ella representa a cuatro de los datos

originales en la señal.

Al cambiar la óptica hacia un extremo opuesto, es completamente natural pensar en un proceso ideal como el que aplique una transformación basada en Wavelets de alta escala, donde lo que se encuentre como resultado son rasgos de baja frecuencia en los que la caracterización se lleve a cabo mediante un volumen de datos pequeño. Después de todo, en lo que respecta a la comodidad en el manejo de los datos es mucho mejor si cada elemento en el vector de características representa 8, 16, 32, 64, 128 o más de los datos en la señal original que solamente a 2 o 4. En la práctica, se encuentra que hay que buscar un equilibrio entre el costo y la ganancia y, en este caso, no se puede continuar elevando la escala de manera indiscriminada sin perder calidad en la información restante desde el punto de vista del reconocimiento. La tendencia generalizada que se revela en el estudio de las curvas ROC es la de que los clasificadores que emplean Wavelets de escala 5 o descomposición en escalas aún más profundas se alejan de la esquina superior izquierda del gráfico irremediablemente.

Considérese que se efectúa una transformación con Wavelets en escala 5 para construir un sistema de reconocimiento de voz. Los indicios que se colectaron apuntan hacia el hecho de que la naturaleza del Wavelet empleado se convierte ahora en un factor importante a considerar, y que, para los efectos de esta investigación el mejor resultado al utilizar una transformación Wavelet de escala 5 se obtendrá de la aplicación del Wavelet de Daubechies de orden 3 antes que con cualquiera de los restantes que se probaron. No debe perderse de vista que esta afirmación es sólo válida para estos escala y orden de Wavelet, dado que para este mismo orden, la elección de una escala menor, digamos 4 o 3 podría producir un clasificador con un desempeño más pobre que el de la combinación antes mencionada. La figura 5.15 muestra diez clasificadores basados en Wavelet de Haar y de Daubechies de ordenes 2 a 10 todos ellos de escala 5; destaca el acercamiento a la esquina superior izquierda de la curva ROC correspondiente al Wavelet de Daubechies de orden 3 db03.

En la búsqueda de un compromiso apropiado entre el volumen de datos que ha de representar a las señales (lo cual se intenta mantener lo más reducido posible) y la cantidad de características de alta frecuencia desechadas (la cual se trata de llevar hasta su límite alto), tratando a la vez de mantener rasgos distintivos de baja frecuencia útiles para efectuar

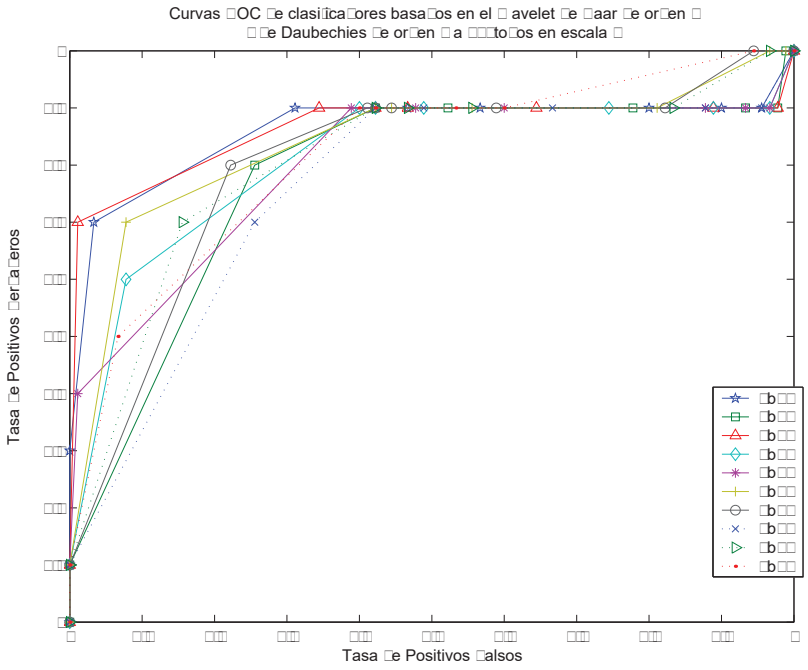

Figura 5.15: Curvas ROC para diez clasificadores basados en los wavelets de Haar y de Daubechies de ordenes 2 a 10, todos ellos en escala 5.

la tarea de reconocimiento (de los cuales se intenta conservar la más alta calidad disponible), se encuentra que una escala pertinente de aplicación de los Wavelets podría bien ser la cuarta. En un nivel como tal, 16 de los datos en la señal original se hallan representados por cada uno de los elementos del vector de características. Hablando en términos de compresión de datos, se trataría de una tasa de 16:1, lo que para muchas aplicaciones puede resultar suficientemente bueno. Examinando los 10 distintos clasificadores construidos en base a Wavelets en esta escala, destaca que los mejores de ellos son, en orden de utilidad, los que emplean: a) el Wavelet de Haar *db01*, b) el Wavelet de Daubechies de orden 4 *db04*, c) el Wavelet *db10*; esto puede apreciarse en la Figura 5.16. Es notorio el hecho de que el Wavelet *db03* que demostró propiedades deseables desde el punto de vista del reconocimiento cuando se empleaba escala 5, constituye ahora en escala 4 la peor de entre las alternativas estudiadas al observarse su extremo alejamiento de la esquina superior izquierda del gráfico.

La tendencia en la evolución de los recursos de cómputo hace pensar en la posibilidad de contar en el corto plazo con máquinas más poderosas y menos costosas que en el

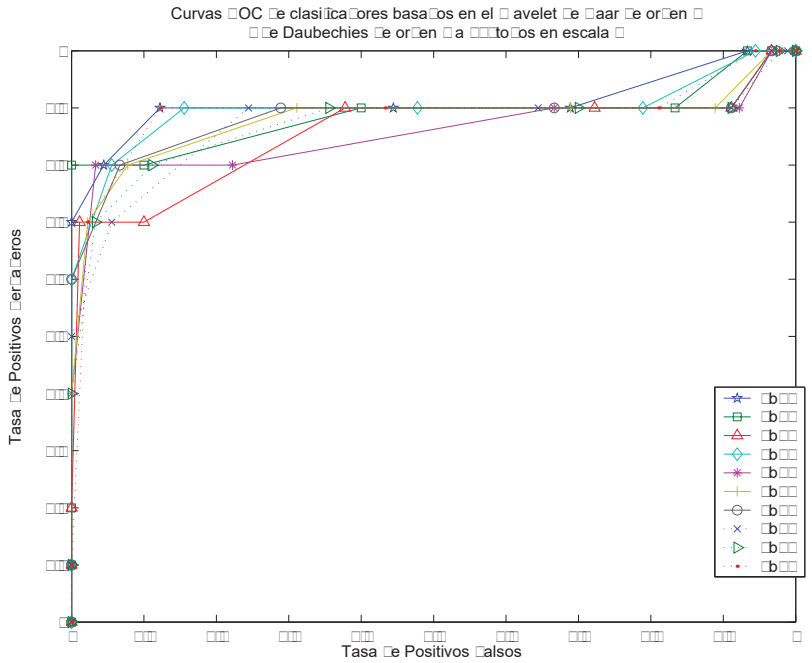

Figura 5.16: Curvas ROC para tres clasificadores basados en el Wavelet de Daubechies de escala 4 y además un clasificador basado en el Wavelet de Daubechies de escala 3.

momento presente. Ello puede permitir al reconocimiento basado en Wavelets el empleo de escalas más bajas de transformación que a pesar de formular vectores de características más largos, sean manipulables en tiempos breves. En un escenario como este conviene entonces examinar el desempeño de los clasificadores que aprovechan Wavelets de escala 3 (una entre los grupos de escalas 1 con 2, que ya se sabe, tendrán los mejores resultados y escalas 4 y 5 que comienzan a resentir la presencia de información relevante para el reconocimiento). Las curvas ROC en la Figura 5.17 revelan que solamente el Wavelet *db10* en escala 3 supera los Wavelets *db01*, *db04* y *db10* de escala 4. Aún cuando también otros clasificadores que incorporan Wavelets de escala 3 superan (en algunos casos ampliamente, incluso) a clasificadores con Wavelets de escala 4, no se trata de los mejores en esta última escala, por lo que la comparación resulta ociosa.

Ha resultado complicado encontrar una manera de cotejar los resultados aquí expresados con los que reportan otros investigadores, dada la peculiaridad de la estructura de la solución a los problemas que se propone. Por ejemplo, [Guevara07] utiliza Wavelets para

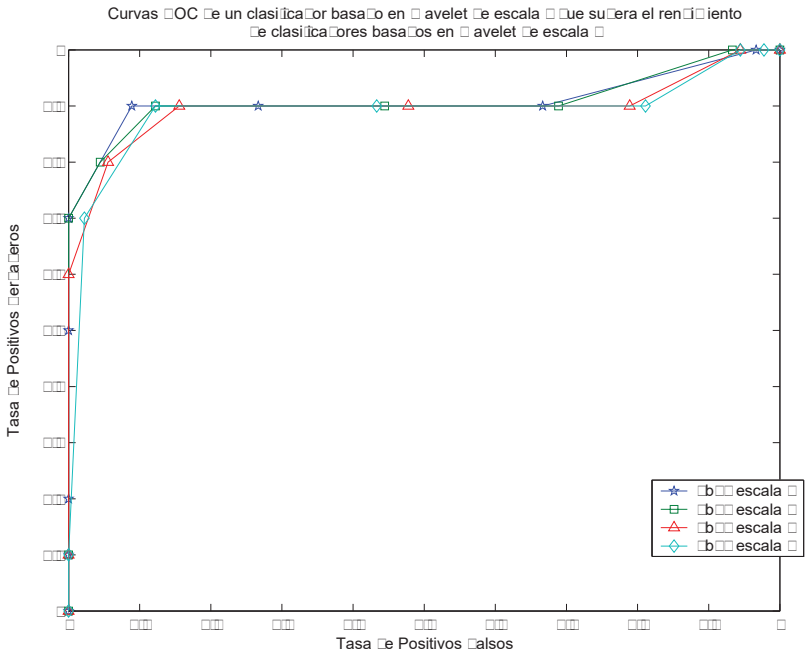

Figura 5.17: Curvas ROC para diez clasificadores basados en los wavelets de Haar y de Daubechies de ordenes 2 a 10, todos ellos en escala 4.

extraer características de la señal de voz, pero aplica ventanas para segmentar esta antes de proceder al cálculo de coeficientes, responsabilidad que en este trabajo se le concede al soporte compacto de las funciones base Wavelet y a su parámetro de traslación. Otra característica de tal sistema es que efectúa descomposición en escala 7 cuando aprovecha el Wavelet de Haar, mientras que descompone hasta escala 6 para un Wavelet de Daubechies y en escala 5 para otro de los miembros de la familia de Wavelets de Daubechies. Esa propuesta alternativa sigue el estilo de los reconocedores basados en MFCC (Mel Frequency Cepstral Coefficients), esencialmente sustituyendo la STFT con DWT. Del mismo modo, en el reconocimiento emplea distancia de Chebyshev y en la evaluación del desempeño del reconocedor utiliza un método estadístico que denomina: Prueba JI Cuadrado de MC Nemar-Datos Correlacionados y no curvas ROC. Este último pareciera ser el componente imparcial que permite hacer converger distintos esquemas de solución al problema para que se les pueda contrastar. Este tipo de dificultad para encontrar sistemas comparables al que se propone siembra la duda respecto de si en realidad es pertinente contrastar elementos que quizá en el fondo sean en alguna cierta medida intrínsecamente incomparables y

posiblemente desvirtuar los logros o bondades de cada propuesta.

El informe de [Guevara07] menciona que respecto del método Coeficientes Cepstrales en Escala Mel se encuentra una tasa de aceptación de 85.32 % y un error de 14.68 % (suma de 100 %). En cuanto al uso de Wavelet de Haar se tiene una tasa de aceptación de 34.47 % y un error de 65.53 % (suma de 100 %). Por lo que respecta a uso del Wavelet de Daubechies 4 (4 por el número de coeficientes en el filtro, lo que corresponde a Daubechies de 2^o orden, db02, en el presente trabajo) se obtiene una tasa de aceptación de 51.79 % y un error de 48.21 % (suma de 100 %). Finalmente, empleando el Wavelet de Daubechies 6 (Daubechies de 3^o orden, db03) se encontró una tasa de aceptación de 61.32 % y un error de 38.68 % (suma de 100 %). Todo hace suponer que la única conclusión común que se puede externar de una comparación entre ambos estudios es el hecho de que el tipo de función base Wavelet elegida impactará la eficiencia de reconocimiento. Algo que parece ser consistente entre ambos informes es el hecho de que [Guevara07] reporta un rendimiento magro cuando utiliza Wavelet de Haar. Esto puede deberse a la profundidad del análisis que se deja llegar hasta la escala 7. El presente trabajo evidencía el hecho de que a partir de la quinta escala se puede esperar un comportamiento decadente de un reconocedor, independientemente del tipo de Wavelet en el que se base.

Capítulo 6

Conclusiones

6.1. Conclusiones Generales

El campo del reconocimiento de voz se encuentra sometido a un intenso desarrollo, dada la activa búsqueda de nuevas pero sobre todo mejores soluciones a cada uno de los múltiples subproblemas que el reconocer/comprender los elementos del lenguaje humano implica. Aún con los enormes avances recientes y la creciente cantidad de especialistas en el área, los principales problemas asociados al reconocimiento de voz permanecen abiertos.

El diseño de sistemas de reconocimiento de voz de calidad requiere la aplicación de técnicas y métodos eficaces en todas las etapas que lo componen. Si alguno de los módulos del sistema opera deficientemente, puede opacar el buen desempeño que otros módulos pueden tener, y en muchas ocasiones provocan que se rompa el frágil equilibrio que hace a un sistema estable y utilizable o bien deficiente e incapaz de brindar el servicio de él demandado.

Las herramientas matemáticas de uso reciente en el campo del procesamiento de señales, lejos de reemplazar a las técnicas tradicionales, deben considerarse como una alternativa que hace más amplio y rico el espectro de metodologías de análisis de que se dispone. Es de esperarse, por lo tanto, que la preocupación central de los analistas siga siendo al menos por un tiempo el respecto de la elección de las herramientas idóneas y su utilización en la forma más conveniente de acuerdo con las condiciones de cada problema en particular.

La implantación en código de los algoritmos que atacan cada uno de los subproblemas en el reconocimiento de voz, abre un inmenso horizonte a la investigación y la experimentación. Del mismo modo, el disponer de sistemas de reconocimiento diseñados para su uso en el laboratorio constituye una herramienta invaluable para la enseñanza y el aprendizaje.

6.2. Conclusiones Específicas

1. Las gráficas ROC de todos los 55 clasificadores considerados en este proyecto (11 diferentes wavelets, cada uno aplicado en 5 niveles de escala) demuestran que cualquiera de ellos funciona de manera aceptable, y que los que presentan mejor rendimiento se acercan bastante en eficiencia a sistemas de reconocimiento automático basados en estrategias similares a la que se presenta en este trabajo [Guevara07].

2. De entre los Wavelets explorados, el que permite la construcción del mejor sistema de reconocimiento de palabras aisladas con una transformación de escala 5 es el de Daubechies de orden 3 (*db03*).

3. De entre los Wavelets explorados, el que permite la construcción del mejor sistema de reconocimiento de palabras aisladas con una transformación de escala 4 es el de Haar (*db01*).

4. Al inicio de la investigación se tenían expectativas limitadas acerca del rendimiento que un sistema de reconocimiento de voz basado en el wavelet de Haar podría llegar a alcanzar, sobre todo debido a lo que se encuentra en la bibliografía como limitaciones de tal función base, entre las cuales se destacaba la imposibilidad de efectuar diferenciación continua. Los resultados indican que si bien este tipo de wavelet podría estar acotado en su aplicación para otros campos, es una herramienta útil en el reconocimiento de palabras aisladas.

5. Como se anticipaba, en términos generales el rendimiento de un clasificador basado en un wavelet de escala 1 (donde la caracterización se efectúa con una cantidad de datos que representa la mitad de los presentes en la señal original), es más alto que el

rendimiento de otro clasificador basado en el mismo wavelet pero este último de escala 5 (donde la caracterización se efectúa con datos que cada uno representa a 32 de los datos presentes en la señal original). La tendencia generalizada que se observa es la aparición de la curva correspondiente al primero de estos clasificadores al "noroeste" de la curva que le corresponde al segundo de ellos, para una parte significativa de ambos trazos.

6.3. Trabajos Futuros

Existen algunas fases de la construcción de un sistema de reconocimiento de voz completo, que no han sido exploradas en esta investigación. Entre ellas puede mencionarse:

1. Aún cuando la segmentación utilizando el criterio de Régimen de Cruces por Cero por sí solo y también en combinación con el cálculo de la Energía de Tiempo Corto funcionó adecuadamente para el sistema que se implantó, posiblemente una combinación de hardware distinto afecte al sistema si no se eliminan muestras indeseadas antes y después de las palabras, de tal manera, es recomendable explorar alguna(s) otra(s) combinación(es) de criterios adicionales, entre los que se puede mencionar el de Magnitud Promedio de Tiempo Corto.

2. Solamente se ha explorado la utilización de un par de Wavelets diferentes, de los que se utilizan directamente los valores de coeficientes que definen el filtro pasabajas en expresiones desarrolladas, es decir, no se implantaron las operaciones matriciales como se mencionaron en la Sección 2.3.3.1, sino las operaciones indicadas en la línea 5 de Algoritmo 1 y en las líneas 12, 13 de Algoritmo 2. Es conveniente que se realice un análisis concienzudo de la complejidad de los algoritmos y juzgar si puede resultar conveniente la implementación de Transformada con Wavelets que opere basada en matrices de transformación donde se alojen la creciente cantidad de coeficientes de definición de Wavelets a medida que se incrementa el orden de los Wavelets de Daubechies, por ejemplo.

3. La construcción automática de diccionarios con el sistema, donde el usuario pueda

elegir el diccionario y se anexe al mismo la caracterización de una palabra que pronuncia.

4. Es posible imaginar distintas formas de interpretación de los resultados al aplicar una transformación entre el dominio del tiempo y el dominio de tiempo-escala que puedan beneficiar al proceso de reconocimiento de voz, por ejemplo, pueden combinarse varios vectores de características de baja frecuencia en sus escalas correspondientes, o bien seguir un esquema típico del problema de codificación de señales, donde además del contenido de baja frecuencia participa también algo del contenido de alta frecuencia de la señal. Esto es, puede continuarse desechando la parte de alta frecuencia en las señales, incorporarla parcialmente, o considerarla por completo para generar un árbol íntegro de descomposición, del cual se puedan seleccionar gran diversidad de distintas combinaciones de vectores de características en cada caso de estudio en particular.

Apéndice A

Construcción de un Wavelet

La elección de los coeficientes de los filtros del Wavelet determina la apariencia del Wavelet que se usa para efectuar el análisis. De hecho, para la construcción de un Wavelet de alguna utilidad práctica no se comienza trazando una forma de onda. En lugar de ello, usualmente tiene más sentido el diseñar los filtros de espejo en cuadratura y luego usarlos para crear la forma de onda. El ejemplo siguiente muestra como se hace esto para el Wavelet de Daubechies de segundo orden.

A.1. Primera iteración

En primera instancia se consideran los coeficientes del filtro pasabajas de reconstrucción L' para el Wavelet db02.

>> **Lprimo=[0.3415 0.5915 0.1585 -0.0915]**

Lprimo =

0.3415 0.5915 0.1585 -0.0915

A continuación los elementos del filtro se colocan en reversa y se multiplica cada segunda muestra por -1 para formar H', el cual es el filtro pasaaltas de reconstrucción.

>> **Hprimo=fliplr(Lprimo); for i=2:2:length(Hprimo) Hprimo(i)=-Hprimo(i); end; Hprimo**

```
Hprimo =
   -0.0915   -0.1585   0.5915   -0.3415
```

Enseguida se sobremuestrea H' por dos mediante la inserción de ceros en sus posiciones alternas, y con ello se obtiene HU.

```
>> HU=[Hprimo(1)]; for i=2:length(Hprimo) HU=[HU 0 Hprimo(i)]; end; HU
HU =
   -0.0915   0   -0.1585   0   0.5915   0   -0.3415
```

Finalmente, se convoluciona el vector sobremuestreado con el filtro pasabajas original para obtener $H2$, el cual alojará la forma de onda del Wavelet que se puede observar en la Figura A.1.

```
>> H2=conv(HU,Lprimo)
H2 =
Columns 1 through 9
   -0.0312   -0.0541   -0.0686   -0.0854   0.1769   0.3644   -0.0229   -0.2561   -0.0541
Column 10
   0.0312
>> plot(H2)
```

A.2. Segunda iteración

Para efectuar los siguientes pasos de cada una de las iteraciones en el proceso de repetición para la generación de la forma de onda del Wavelet de Daubechies de segundo orden, se sobremuestrea el vector $H2$ resultante en la iteración previa para encontrar un HU actualizado, y se convoluciona con el filtro pasabajas original L' con lo que el nuevo $H2$ aloja la forma de onda del Wavelet en la iteración presente, la cual se presenta en la Figura A.2 de cuyo examen resulta evidente que comienza a revelarse un patrón.

```
>> HU=[H2(1)]; for i=2:length(H2) HU=[HU 0 H2(i)]; end; HU
```

```
HU =
Columns 1 through 9
   -0.0312    0   -0.0541    0   -0.0686    0   -0.0854    0    0.1769
Columns 10 through 18
    0    0.3644    0   -0.0229    0   -0.2561    0   -0.0541    0
Column 19
    0.0312
>> H2=conv(HU,Lprimo)
H2 =
Columns 1 through 9
   -0.0107   -0.0185   -0.0234   -0.0292   -0.0320   -0.0356   -0.0400   -0.0442    0.0469
Columns 10 through 18
    0.1124    0.1525    0.1993    0.0499   -0.0469   -0.0911   -0.1494   -0.0591   -0.0086
Columns 19 through 22
    0.0021    0.0234    0.0050   -0.0029
>> plot(H2)
```

A.3. Tercera iteración

La Figura A.3 muestra la forma de onda de Wavelet de Daubechies de segundo orden en tres iteraciones, tras la ejecución de los dos pasos que se indican a continuación.

```
>> HU=[H2(1)]; for i=2:length(H2) HU=[HU 0 H2(i)]; end; HU
HU =
Columns 1 through 9
   -0.0107    0   -0.0185    0   -0.0234    0   -0.0292    0   -0.0320
Columns 10 through 18
    0   -0.0356    0   -0.0400    0   -0.0442    0    0.0469    0
Columns 19 through 27
    0.1124    0    0.1525    0    0.1993    0    0.0499    0   -0.0469
```

```
Columns 28 through 36
   0         -0.0911    0         -0.1494    0         -0.0591    0         -0.0086    0
Columns 37 through 43
   0.0021    0          0.0234    0          0.0050    0          -0.0029
>> H2=conv(HU,Lprimo)
H2 =
Columns 1 through 9
   -0.0036   -0.0063   -0.0080   -0.0100   -0.0109   -0.0122   -0.0137   -0.0151   -0.0156
Columns 10 through 18
   -0.0163   -0.0172   -0.0182   -0.0193   -0.0204   -0.0214   -0.0225   0.0090    0.0318
Columns 19 through 27
   0.0458    0.0622    0.0699    0.0799    0.0922    0.1040    0.0487    0.0113    -0.0081
Columns 28 through 36
   -0.0323   -0.0385   -0.0496   -0.0655   -0.0800   -0.0439   -0.0213   -0.0123   0.0003
Columns 37 through 45
   -0.0006   0.0020    0.0083    0.0137    0.0054    0.0008    -0.0002   -0.0021   -0.0005
Column 46
   0.0003
>> plot(H2)
```

A.4. Cuarta iteración

La Figura A.4 muestra la forma de onda de Wavelet de Daubechies de segundo orden en cuatro iteraciones, tras la ejecución de los dos pasos que se indican a continuación.

```
>> HU=[H2(1)]; for i=2:length(H2) HU=[HU 0 H2(i)]; end; HU
HU =
Columns 1 through 9
   -0.0036   0         -0.0063   0         -0.0080   0         -0.0100   0         -0.0109
Columns 10 through 18
```

```
   0         -0.0122    0         -0.0137    0         -0.0151    0         -0.0156    0
Columns 19 through 27
   -0.0163    0         -0.0172    0         -0.0182    0         -0.0193    0         -0.0204
Columns 28 through 36
   0         -0.0214    0         -0.0225    0         0.0090     0         0.0318     0
Columns 37 through 45
   0.0458     0         0.0622     0         0.0699     0         0.0799     0         0.0922
Columns 46 through 54
   0         0.1040     0         0.0487     0         0.0113     0         -0.0081    0
Columns 55 through 63
   -0.0323    0         -0.0385    0         -0.0496    0         -0.0655    0         -0.0800
Columns 64 through 72
   0         -0.0439    0         -0.0213    0         -0.0123    0         0.0003     0
Columns 73 through 81
   -0.0006    0         0.0020     0         0.0083     0         0.0137     0         0.0054
Columns 82 through 90
   0         0.0008     0         -0.0002    0         -0.0021    0         -0.0005    0
Column 91
   0.0003
>> H2=conv(HU,Lprimo)
H2 =
Columns 1 through 9
   -0.0012   -0.0022   -0.0027   -0.0034   -0.0037   -0.0042   -0.0047   -0.0052   -0.0053
Columns 10 through 18
   -0.0056   -0.0059   -0.0062   -0.0066   -0.0070   -0.0073   -0.0077   -0.0077   -0.0078
Columns 19 through 27
   -0.0080   -0.0082   -0.0085   -0.0087   -0.0089   -0.0092   -0.0095   -0.0098   -0.0100
Columns 28 through 36
   -0.0103   -0.0106   -0.0108   -0.0111   -0.0113   -0.0005    0.0074    0.0123    0.0180
```

```
Columns 37 through 45
   0.0207    0.0242    0.0285    0.0326    0.0337    0.0356    0.0384    0.0409    0.0442
Columns 46 through 54
   0.0473    0.0501    0.0531    0.0331    0.0193    0.0116    0.0022   -0.0010   -0.0058
Columns 55 through 63
  -0.0123   -0.0184   -0.0183   -0.0198   -0.0230   -0.0258   -0.0302   -0.0342   -0.0377
Columns 64 through 72
  -0.0414   -0.0277   -0.0186   -0.0142   -0.0086   -0.0076   -0.0053   -0.0018    0.0013
Columns 73 through 81
  -0.0002   -0.0004    0.0006    0.0013    0.0032    0.0047    0.0060    0.0073    0.0040
Columns 82 through 90
   0.0019    0.0011   -0.0000    0.0001   -0.0002   -0.0008   -0.0013   -0.0005   -0.0001
Columns 91 through 94
   0.0000    0.0002    0.0000   -0.0000
>> plot(H2)
```

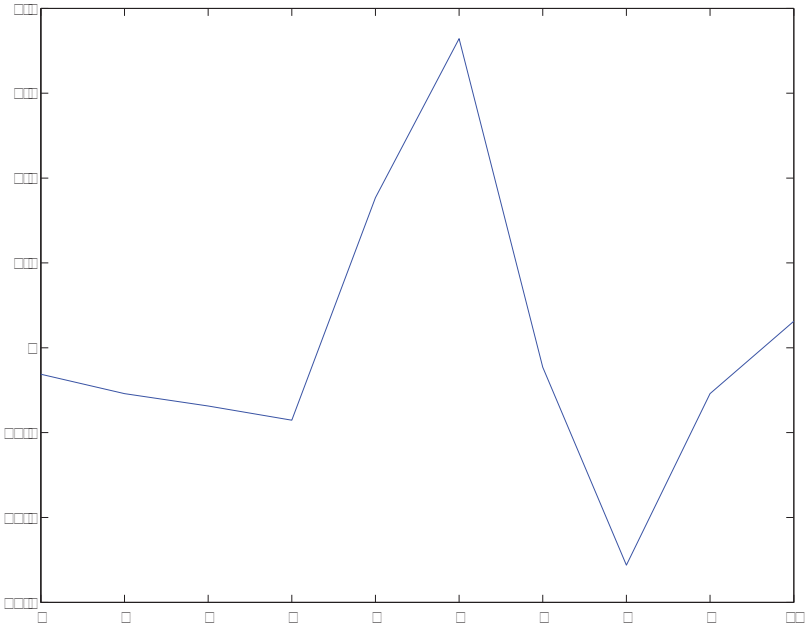

Figura A.1: Primera iteración para la generación de la forma de onda del Wavelet db02.

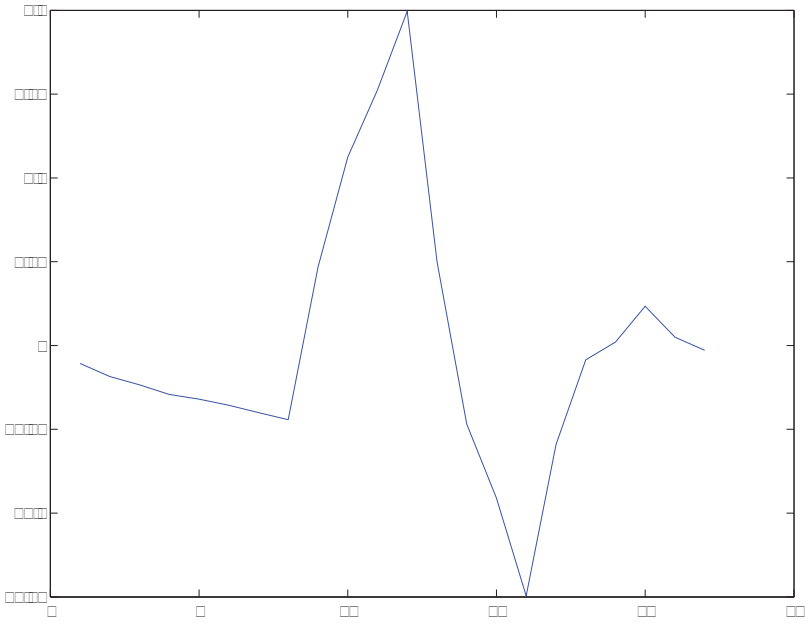

Figura A.2: Segunda iteración para la generación de la forma de onda del Wavelet db02.

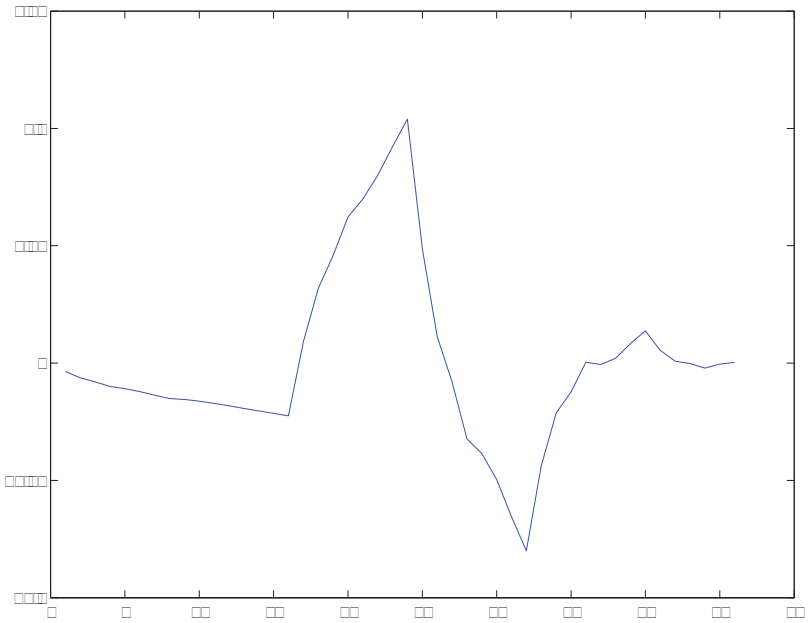

Figura A.3: Tercera iteración para la generación de la forma de onda del Wavelet db02.

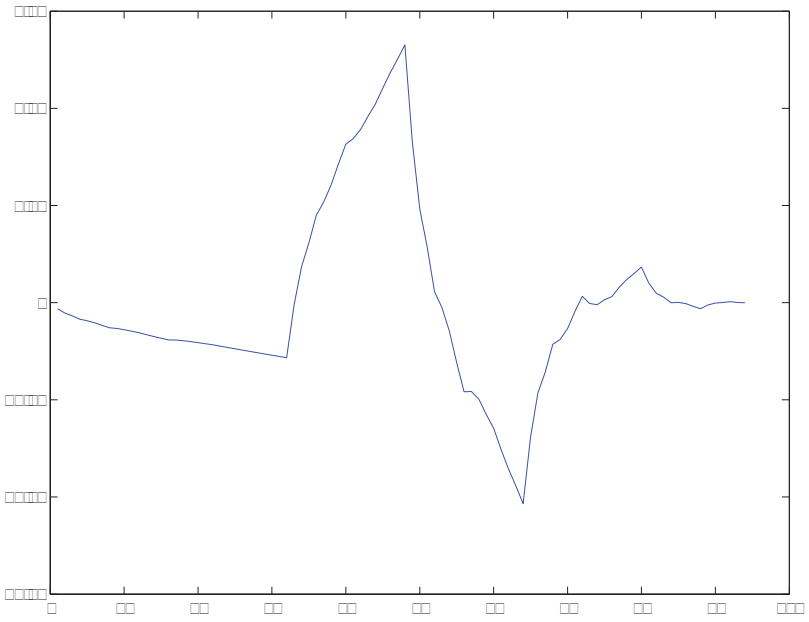

Figura A.4: Cuarta iteración para la generación de la forma de onda del Wavelet db02.

Referencias

[CMUISL07] CMUISL, I. C. The JANUS speech translation system. http://www.is.cs.cmu.edu/mie/janus.html, 1999–2007.
URL `http://www.is.cs.cmu.edu/mie/janus.html`

[CMULPG07] CMULPG, C. C. Project LISTEN. A Reading Tutor that Listens. http://www.cs.cmu.edu/~ listen/index.html, 1996–2007.
URL `http://www.cs.cmu.edu/ listen/index.html`

[CMURSRG07] CMURSRG, C. C. The CMU Sphinx Group Open Source Speech Recognition Engines. http://cmusphinx.sourceforge.net/html/cmusphinx.php, 1996–2007.
URL `http://cmusphinx.sourceforge.net/html/cmusphinx.php`

[Cohen96] Cohen, A. y Kovacevic, J. Wavelets: The mathematical background. tomo 84 de *Proceedings of the IEEE: Special issue on Wavelets*, págs. 514–522. The Institute of Electrical and Electronics Engineers, Inc., April 1996. ISSN 0018-9219.

[Daubechies96] Daubechies, I. Where do wavelets come from? - a personal point of view. tomo 84 de *Proceedings of the IEEE: Special issue on Wavelets*, págs. 510–513. The Institute of Electrical and Electronics Engineers, Inc., April 1996. ISSN 0018-9219.

[Fawcett04] Fawcett, T. Roc graphs: Notes and practical considerations for researchers. 2004.

[Gillemain96] Gillemain, P. y Kronland-Martinet, R. Characterization of acoustic signals through continuous linear time-frequency representations. tomo 84 de *Proceedings of the IEEE: Special issue on Wavelets*, págs. 561–585. The Institute of Electrical and Electronics Engineers, Inc., April 1996. ISSN 0018-9219.

[Graps95] Graps, A. An introduction to wavelets. tomo 2 de *IEEE Computational Science and Engineering*. The Institute of Electrical and Electronics Engineers, Inc., 1995.

[Guevara07] Guevara, J. L. y Salazar, J. O. Extracción de características en el procesamiento digital de una señal para el mejoramiento del reconocimiento automático de habla usando wavelets. 2007.

[Hansen97] Hansen, C. H. y Snyder, S. D. *Active Control of Noise and Vibration*, págs. 200–204. Taylor and Francis, 1ª ed$^{\underline{\text{ón}}}$., 1997. ISBN 0419193901.

[Huang92] Huang, X., Alleva, F., Hon, H.-W., Hwang, M.-Y., y Rosenfeld, R. The sphinx-ii speech recognition system: An overview. 1992.

[Janer96] Janer, L., Martí, J., Nadeu, C., y Leida-Solano, E. Wavelet transforms for non-uniform speech recognition systems. tomo 4 de *IEEE Sixth International Conference on Spoken Language Processing*, págs. 2348–2351. Dept. de Teoria del Senyal i Comunicacions, Univ. Politecnica de Catalunya, Barcelona, The Institute of Electrical and Electronics Engineers, Inc., October 1996.

[Lea79] Lea, W. A. y Shoup, J. E. Review of the arpa sur project and survey of current technology in speech understanding. Inf. téc., January 1979.

[Lee90] Lee, K.-F., Hon, H.-W., y Reddy, R. An overview of the sphinx speech recognition system. IEEE Transactions On Acoustics, Speech And Signal Processing, págs. 35–45. The Institute of Electrical and Electronics Engineers, Inc., January 1990.

[Lieberman88] Lieberman, P. y Blumstein, S. E. *Speech Physiology, Speech Perception, and Acoustic Phonetics*, págs. 16–33. Cambridge Studies in Speech Science and Communication, 1ª edón., 1988. ISBN 0521308666.

[Mariani89] Mariani, J. Recent advances in speech processing. IEEE International Conference on Acoustics, Speech, and Signal Processing. The Institute of Electrical and Electronics Engineers, Inc., 1989.

[Meyer93] Meyer, Y. *Wavelets - Algorithms and Applications*, págs. 1–73. Society for Industrial and Applied Mathematics, 1ª edón., 1993. ISBN 0-89871-309-9.

[Misiti96] Misiti, M., Misiti, Y., Oppenheim, G., y Poggi, J.-M. *Wavelet Toolbox For Use With Matlab*, págs. (1–1)–(1–36), (6–1)–(6–73). The Math Works, Inc., 1ª edón., 1996.

[Myers81] Myers, C. S., Rabiner, L. R., y Rosenberg, A. E. On the use of dynamic time warping for word spotting and connected word recognition. tomo 60 de *The Bell System Technical Journal*, págs. 303–311. March 1981. ISSN 0005-8580.

[Pellom00] Pellom, B., Ward, W., y Pradhan, S. The cu communicator: An architecture for dialogue systems. IEEE Sixth International Conference on Spoken Language Processing. Center for Spoken Language Research, University of Colorado, Boulder, CO, USA, The Institute of Electrical and Electronics Engineers, Inc., 2000.

[Proakis92] Proakis, J. G. y Manolakis, D. G. *Digital Signal Processing - Principles, Algorithms and Applications*, págs. 1–51, 143–213, 395–456, 684–735. Macmillan Publishing Company, 1ª edón., 1992. ISBN 0-02-396815-X.

[Rabiner78] Rabiner, L. R. y Schafer, R. W. *Digital Processing of Speech Signals*. Prentice Hall, 1ª edón., 1978. ISBN 0-13-213603-1.

[Reddy76] Reddy, D., Erman, L., Fennell, R., y Neely, R. The hearsay speech understanding system: An example of the recognition process. tomo C-25

de *IEEE Transactions on Computers*, págs. 422–431. Department of Computer Science, Carnegie-Mellon University, The Institute of Electrical and Electronics Engineers, Inc., April 1976.

[Reddy89] Reddy, R. Speech research at carnegie mellon. *En HLT '89: Proceedings of the workshop on Speech and Natural Language*, págs. 119–119. Association for Computational Linguistics, Morristown, NJ, USA, 1989. doi: http://dx.doi.org/10.3115/100964.1138536.

[Smith97] Smith, S. W. *The Scientist and Engineer's Guide to Digital Signal Processing*, págs. 169–184, 285–296. California Technical Pub., 1ª ed$^{\underline{\text{ón}}}$., 1997. ISBN 978-0966017632.

[Spackman89] Spackman, K. A. Signal detection theory: valuable tools for evaluating inductive learning. *En Proceedings of the sixth international workshop on Machine learning*, págs. 160–163. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1989.

[Strang97] Strang, G. y Nguyen, T. *Wavelets and Filter Banks*, págs. 1–102, 174–218. Wellesley-Cambridge Press, 1ª ed$^{\underline{\text{ón}}}$., 1997. ISBN 0-9614088-7-1.

[UCCSLR07] UCCSLR, T. U. CU Communicator Spoken Dialog System. http://cslr.colorado.edu/beginweb/cumove/cucommunicator.html, 1999–2007.
URL `http://cslr.colorado.edu/beginweb/cumove/cucommunicator.html`

[Waibel96] Waibel, A. Interactive translation of conversational speech. 29:41–48, July 1996.

[Ward91] Ward, W. H. The phoenix system: Understanding spontaneous speech. IEEE International Conference on Acoustics, Speech, and Signal Processing. The Institute of Electrical and Electronics Engineers, Inc., April 1991.

[Ward99] Ward, W. y Pellom, B. The cu communicator system. IEEE Workshop

On Automatic Speech Recognition and Understanding. The Institute of Electrical and Electronics Engineers, Inc., 1999.

[Wesfreid93] Wesfreid, E. y Wickerhauser, M. V. Adapted local trigonometric transforms and speech processing. 41, December 1993.

Glosario

ARPA *Advanced Research Projects Agency*-Agencia de Proyectos de Investigación Avanzados

CFT *Continuous Fourier Transform*-Transformada de Fourier Contínua

CWT *Continuous Wavelet Transform*-Transformada Wavelet Contínua

DARPA *Defense Advanced Research Projects Agency*-Agencia de Proyectos de Investigación Avanzados de Defensa

Daubechies *Ingrid Daubechies* (17 de Agosto de 1954 -), Matemática y Física Belga. Ver nota al pie en la página 41.

DFT *Discrete Fourier Transform*-Transformada Discreta de Fourier

DTW *Dynamic Time Warping*-Doblado Dinámico en Tiempo

DWT *Discrete Wavelet Transform*-Transformada Wavelet Discreta

Haar *Alfrèd Haar* (11 de Octubre de 1885 - 16 de Marzo de 1933), Matemático Húngaro. Ver nota al pie en la página 39.

IBM *International Business Machines*-Máquinas de Negocios Internacionales

LISTEN *Literacy Innovation that Speech Technology Enables*-Inovación en Alfabetización que Hace Posible la Tecnología del Habla

Mallat *Stèphane Georges Mallat* Matemático Francés. Ver nota al pie en la página 34.

NIST *National Institute of Standards and Technology*-Instituto Nacional de Estándares y tecnología

NSF *National Science Foundation*-Fundación Nacional de la Ciencia

PCM *Pulse Code Modulation*-Modulación por Código de Impulsos

ROC *Receiver Operating Characteristic*-Característica de Operación de Receptor

SAM *Speech Assessment Methodology*-Metodología de Evaluación del Habla

SDT *Signal Detection Theory*-Teoría de Detección de Señales

STFT *Short Time Fourier Transform*-Transformada de Fourier de Tiempo Corto