

UNIVERSIDAD NACIONAL DE TRUJILLO
DEPARTAMENTO DE CIENCIAS FÍSICAS Y MATEMÁTICAS
ESCUELA DE INFORMÁTICA

**EXTRACCIÓN DE CARACTERÍSTICAS EN EL
PROCESAMIENTO DIGITAL DE UNA SEÑAL PARA EL
MEJORAMIENTO DEL RECONOCIMIENTO
AUTOMÁTICO DEL HABLA USANDO WAVELETS**

por

Jorge Luis Guevara Díaz Juan Orlando Salazar Campos

TESIS PRESENTADA
PARA OBTENER EL TÍTULO PROFESIONAL DE:
INGENIERO INFORMÁTICO

JURADO DE TESIS:

Carlos Castillo Diestra
Msc. Sistemas de Información

Ronald León Navarro
Msc. Matemática Industrial

Ely Miguel Aguilar
Msc. Física

Trujillo, Perú

Enero, 2007

RESUMEN

EXTRACCIÓN DE CARACTERÍSTICAS EN EL PROCESAMIENTO DIGITAL DE UNA SEÑAL PARA EL MEJORAMIENTO DEL RECONOCIMIENTO AUTOMÁTICO DEL HABLA USANDO WAVELETS

por

Jorge Luis Guevara Díaz Juan Orlando Salazar Campos

La presente tesis, centra su investigación en la extracción de características de una señal de habla que serán los valores que mejor permitan representarla, los mejores valores de características serán aquellos que contengan mayor información, esta información será usada para hacer un reconocimiento automático de la palabra hablada en el ordenador.

El presente trabajo enfoca su investigación en el Reconocimiento Automático del Habla (Speech Recognition), cuyo amplio campo de investigación abarca desde áreas como procesamiento digital de señales, análisis y diseño de algoritmos, inteligencia artificial, matemáticas, hasta áreas como la lingüística, fonología entre muchas otras. Nuestra investigación se centra en el desarrollo de los algoritmos correspondientes a la etapa del procesamiento digital de la señal de habla, haciendo un estudio de la

aplicación de las wavelets en este campo, y comparándolas con un análisis mediante la Transformada de Fourier. El presente trabajo de investigación brinda una técnica alternativa en la extracción de características en el proceso de Reconocimiento Automático del Habla haciendo uso de las wavelets.

Palabras clave: Muestreo, Wavelets, cepstrum, Mel, Fourier, Dynamic Time Warping, Procesamiento Digital de Señales.

Agradecimientos

Agradecemos a nuestra familia y amigos por el apoyo y soporte brindado, a los profesores que nos inculcaron el deseo por la investigación y la superación, y a todos los que hicieron posible que esta tesis pudiera realizarse.

A nuestra familia y amigos.

INDICE

Resumen	ii
Agradecimientos	iv
Lista de Figuras	xi
Lista de Tablas	xix
Introducción	xxi
I Marco Teórico	1
1 Sistemas Informáticos del Lenguaje Hablado	2
1.1 Arquitectura	5
1.1.1 Reconocimiento Automático del Habla	5
1.1.2 Síntesis del Habla	6
1.1.3 Entendimiento del Lenguaje Hablado	7
1.2 Trabajos Previos	9
1.3 Historia	11
2 El Habla, Producción y Percepción	14
2.1 El Habla como Sonido	14
2.2 Producción del Habla	16
2.3 Percepción del Habla	17

2.3.1	Fisiología del Oído	19
3	Procesamiento Digital de la Señal	22
3.1	Señales Digitales	22
3.2	Sistemas Digitales	24
3.3	Transformada de Fourier	25
3.3.1	Transformada Discreta de Fourier	26
3.3.2	Complejidad Computacional de la Transformada Discreta de Fourier	26
3.3.3	Transformada Rápida de Fourier	27
3.3.4	Algoritmo Radix-2 con Decimación en Frecuencia y reorde- namiento en la salida de bits mezclados	28
3.3.5	Complejidad Computacional de la Transformada Rápida de Fourier	32
3.4	Función Ventana	34
3.4.1	Ventana Rectangular	34
3.4.2	Ventana Generalizada Hamming	35
3.5	Representación de la Señal de Habla	36
3.5.1	Transformada Corta de Fourier	36
3.5.2	Transformada Discreta del Coseno	38
3.5.3	Procesamiento Cepstral	39

3.5.4	Extracción de características basadas en la Transformada de Fourier	40
3.5.5	Coefficientes Cepstrales en Frecuencia Mel	41
3.6	Muestreo de la Señal de Habla	44
3.7	Codificación de la Señal de Habla	45
3.7.1	Codificadores Escalares de Forma de Onda	46
3.7.2	Codificación PCM	47
4	Wavelets	48
4.1	Introducción	48
4.2	Transformada Wavelet	50
4.3	Transformada Wavelet Continua	51
4.4	Comparación de la Transformada de Fourier con la Transformada Wavelet	52
4.5	Transformada Wavelet Discreta	55
4.5.1	Función Escala	57
4.5.2	Función Wavelet	58
4.6	Análisis Multiresolución	59
4.7	Filtro Pasa Banda	62
4.8	Codificación de Subbanda	64
4.9	Complejidad Computacional de la Transformada Wavelet	67
4.10	Wavelet Packets	69

5	Técnicas para el Reconocimiento Automático del Habla	72
5.1	Redes Neuronales Artificiales	72
5.2	Modelos Ocultos de Markov	73
5.3	Dynamic Time Warping	73
5.3.1	Distancia General Normalizada en el Tiempo	75
5.3.2	Restricciones de la Función Warping	77
5.3.3	Coeficientes de Pesos	80
5.3.4	Algoritmo PD-Matching	82
6	Reconocimiento Automático del Habla utilizando Wavelets	87
6.1	Modelo propuesto para la Extracción de Características basadas en las wavelets	89
6.1.1	Wavelet de Haar	91
6.1.2	Wavelet de Daubechies	92
6.1.3	Wavelet Coiflets	98
6.1.4	Obtención de las Características	101
6.2	Modelo propuesto para la Extracción de Características basadas en las wavelets packet Perceptual con Daubechies 4	101
6.2.1	Obtención de las Características para las Wavelets Packet	103
II	Métodos	108

7	Métodos y Técnicas	109
7.1	Enfoque	109
7.2	Hipótesis	109
7.3	Tipo de Investigación	109
7.4	Universo y Muestra	109
7.5	Instrumentos	110
7.6	Procedimiento	110
7.7	Métodos y procedimientos para la recolección de datos	111
7.8	Análisis estadísticos de los datos	111
III	Resultados y Análisis	112
8	Resultados	113
9	Análisis	136
IV	Conclusiones	144
	Conclusiones	145
	Comentarios	150
	Referencias Bibliográficas	154
	Apéndice	156

Lista de figuras

1.1	Arquitectura básica de un Sistema de Reconocimiento Automático del Habla. Fuente : [Huang and Hon, 2001]	7
1.2	Arquitectura básica de un Sistema de Conversión Texto-Habla. Fuente : [Huang and Hon, 2001]	8
2.1	Diagrama del aparato fonador humano.	18
2.2	Diagrama del sistema auditivo humano. Fuente : Microsoft Encarta 2006.	21
3.1	Representación de un número complejo en el plano cartesiano.	24
3.2	En la parte superior se tiene la señal de habla en el dominio del tiempo. En la parte inferior se tiene la señal de habla en el dominio de la frecuencia.	25
3.3	La mariposa Gentleman-Sande.	31
3.4	La entrada x en el array a es sobrescrita por la salida mezclada X	31
3.5	Algoritmo Radix 2 . Fuente: [Chu and George, 2000]	33
3.6	Ventana Rectangular en el dominio del tiempo y de la frecuencia.	35
3.7	Ventana Hamming en el dominio del tiempo y de la frecuencia.	36

3.8	Comparación en el dominio de la frecuencia entre la Ventana Rectangular y la Ventana Hamming.	37
3.9	Transformada Corta de Fourier.	38
3.10	Modelo básico fuente-filtro para señales de habla.	39
3.11	Coeficientes Cepstrales de la señal de habla, la parte baja corresponde al tracto vocal, la parte alta corresponde a la información provenientes de las cuerdas vocales. Fuente: Oppenheim	40
3.12	Filtros triangulares usados en el cálculo del Mel-Cepstrum.	42
3.13	Escala perceptual Mel comparada con la escala de frecuencias.	43
3.14	Muestreo de una señal de habla.	45
4.1	Ventana Tiempo-Frecuencia para la Transformada de Fourier. Fuente: [Mallat, 1989]	53
4.2	Ventana Tiempo-Frecuencia para la Transformada Wavelet. Fuente: [Mallat, 1989]	54
4.3	Localización de las wavelets discretos en el espacio tiempo-escala en una malla diádica. Fuente: [Mallat, 1989]	56
4.4	Localización de las wavelets discretas en el dominio de la frecuencia, como resultado de escalar las wavelets en el dominio del tiempo.	63
4.5	Función escala en el dominio de la frecuencia y como ésta es analizada por las wavelets.	64

4.6	Banco de Filtros Iterativo.	65
4.7	Esquema del banco de filtros.	67
4.8	Implementación del banco de filtros iterativo.	68
4.9	Árbol de descomposición para el Wavelet Packet, cada nodo forma un espacio W.	69
4.10	Wavelet Packet en profundidad 3, mejor cubrimiento de rango de frecuencias calculado con el Wavelet de Haar.	70
4.11	Wavelet Packet en profundidad 3, mejor cubrimiento de rango de frecuencias calculado con el Wavelet de Daubechies.	70
4.12	Cálculo de las Wavelets Packets mediante banco de filtros iterativo.	71
5.1	Función Warping y Ventana de Ajuste. Fuente: [Sakoe and Chiba, 1978]	77
5.2	Slope Constraint en Función Warping. Fuente: [Sakoe and Chiba, 1978]	79
5.3	Coeficientes de pesos para la forma simétrica y la forma asimétrica. Fuente: [Sakoe and Chiba, 1978]	82
5.4	Diagrama de flujo del algoritmo PD-Matching.	84
5.5	Algoritmos simétricos y asimétricos con condición de Slope Constraint $P = 0, \frac{1}{2}, 1, 2$. Fuente: [Sakoe and Chiba, 1978]	86
6.1	Muestreo de la señal analógica para construir la señal digital.	87
6.2	Obtención de los segmentos con información de la señal de habla.	88

6.3	Función escala ϕ y función wavelet ψ de Haar.	92
6.4	Árbol de descomposición para las wavelets de Haar	93
6.5	Señal de Habla correspondiente a los digitos 10001-90210-01803. . . .	93
6.6	Descomposición de la señal de habla correspondiente a la secuencia de digitos 10001-90210-01803, mediante las wavelets de Haar.	94
6.7	Función escala ϕ y función wavelet ψ Daubechies..	96
6.8	Árbol de descomposición para las wavelets de Daubechies 6	96
6.9	Señal de habla correspondiente a los digitos 10001-90210-01803. . . .	97
6.10	Descomposición de la señal de habla correspondiente a la secuencia de digitos 10001-90210-01803, mediante las wavelets de Daubechies 4 . . .	97
6.11	Descomposición de la señal de habla correspondiente a la secuencia de digitos 10001-90210-01803, mediante las wavelets de Daubechies 6 . . .	98
6.12	Función escala ϕ y función wavelet ψ de Coiflets.	99
6.13	Señal de habla correspondiente a los digitos 10001-90210-01803. . . .	100
6.14	Descomposición de la señal de habla correspondiente a la secuencia de digitos 10001-90210-01803. mediante las wavelets Coiflet 6	100
6.15	Equivalencias de Mel a Frecuencias	102
6.16	Arbol de descomposición espacios de resolución 12,13 y 14	103
6.17	Arbol de descomposición espacios de resolución 9,10 y 11	104
6.18	Arbol de descomposición espacios de resolución 6,7 y 8	105

6.19	Arbol de descomposición espacios de resolución 4 y 5	106
6.20	Arbol de descomposición espacios de resolución 1, 2 y 3	107
8.1	Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets de Haar. Trujillo 2006.	114
8.2	Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets de Daubechies 4. Trujillo 2006.	116
8.3	Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets de Daubechies 6. Trujillo 2006.	117
8.4	Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets de Coiflet 6. Trujillo 2006.	118

8.5 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets Packet Walsh. Trujillo 2006. 119

8.6 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets Packet Daubechies 4. Trujillo 2006. 120

8.7 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets packet Daubechies 6. Trujillo 2006. 121

8.8 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets packet Perceptual con Daubechies 4. Trujillo 2006. 122

8.9 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets de Haar. Trujillo 2006. 123

8.10	Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets de Daubechies 4. Trujillo 2006.	124
8.11	Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets de Daubechies 6. Trujillo 2006.	126
8.12	Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets de Coiflet 6. Trujillo 2006.	127
8.13	Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets Packet Walsh. Trujillo 2006.	128
8.14	Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets Packet Daubechies 4. Trujillo 2006.	129

8.15	Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets packet Daubechies 6. Trujillo 2006.	130
8.16	Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets packet Perceptual con Daubechies 4. Trujillo 2006.	131
9.1	Parte real e imaginaria de las Wavelets de Morlet	151
9.2	Parte real de las Wavelets de Morlet en el dominio del tiempo y en el dominio de la frecuencia en 1D y 2D respectivamente	151
0.3	Lorito graficando la Transformada Discreta de Fourier de una señal de habla.	158
0.4	Lorito graficando el espectograma como paso previo para la extracción de características MFCC.	159
0.5	Lorito mostrando resultados de reconocimiento por medio de las wavelets.	160
0.6	Lorito graficando los coeficientes en diferentes espacios de resolución de una señal de habla	161

Lista de Tablas

1.1	Error humano es aproximadamente 5 veces menor que el error de las máquinas. Fuente : [Huang and Hon, 2001]	13
2.1	Relación entre atributos físicos y perceptuales del sonido. Fuente : [Huang and Hon, 2001]	21
6.1	Rango aproximado de frecuencias en los espacios V y W para una frecuencia de muestreo igual a 16000Hz y un nivel de descomposición $m = 7$	91
8.1	Datos obtenidos utilizando la técnica de DTW como reconocedor, con distancia Euclidiana y con Slope Constrain $P=0$. Se observa la mejor performance en las Wavelet Packet Perceptuales Daubechies 6, y la mas pobre en las Wavelet Haar	132
8.2	Datos obtenidos utilizando la técnica de DTW como reconocedor, con distancia Euclidiana y con Slope Constrain $P=1$. Se observa la mejor performance en las Wavelet Packet Perceptuales Daubechies 6, y la mas pobre en las Wavelet Haar	133

8.3	Datos obtenidos utilizando la técnica de DTW como reconocedor, con distancia Chebyshev y con Slope Constrain $P=0$. Se observa la mejor performance en las Wavelet Packet Perceptuales Daubechies 6, y la mas pobre en las Wavelet Haar	133
8.4	Datos obtenidos utilizando la técnica de DTW como reconocedor, con distancia Chebyshev y con Slope Constrain $P=1$. Se observa la mejor performance en las Wavelet Packet Perceptuales Daubechies 6, y la mas pobre en las Wavelet Haar	134
8.5	Tasa de reconocimiento de las palabras por método.	135
9.1	Datos obtenidos utilizando la técnica de DTW como reconocedor, con distancia Chebyshev y con Slope Constrain $P=1$. Se observa la mejor performance en las Wavelet Continuos de Morlet	152

Introducción

La realidad problemática de los sistemas informáticos actuales de reconocimiento automático del habla, donde está enfocada la mayor parte de la investigación, son principalmente las áreas del procesamiento digital de la señal de habla, donde los científicos de la computación buscan mejores algoritmos para procesar y caracterizar adecuadamente la señal de habla digitalizada en una computadora y el área del reconocimiento donde se buscan también algoritmos y técnicas para que la computadora pueda reconocer adecuadamente la palabra hablada.

La presente tesis hace un estudio de la aplicación de las wavelets en el procesamiento digital de la señal como una alternativa para la extracción de características de una palabra dada; una de las técnicas clásicas más robusta y usada frecuentemente por los reconocedores comerciales en la actualidad es la técnica denominada Coeficientes Cepstrales en Frecuencia Mel, que en la etapa de análisis de las frecuencias presentes en la señal de habla, hace uso de la Transformada de Fourier cuya complejidad computacional para los algoritmos rápidos conocidos es $O(n \log n)$ en el tamaño de la entrada de los datos. En consecuencia la presente investigación se inicia planteando el siguiente problema:

¿El análisis de la señal digital de habla mediante wavelets, proporciona una técnica alternativa, en la extracción de características para el Reconocimiento Automático

del Habla por parte de un ordenador frente a un procesamiento digital de la señal mediante la Transformada de Fourier?.

Pues conociendo que las wavelets tienen mejor localización tiempo-frecuencia en las señales (en este caso señales de habla) que la Transformada de Fourier y también la existencia de algoritmos rápidos con complejidad computacional $O(n)$, nos llevó a plantearnos la siguiente hipótesis de trabajo:

El procesamiento digital de la señal del habla con wavelets nos proporciona una alternativa frente a un análisis de la señal mediante la Transformada de Fourier, mejorando el análisis del espectro, y disminuyendo la complejidad computacional en el análisis de la señal, como paso previo para la obtención de vectores de características en el reconocimiento automático del habla por parte de un ordenador.

Esto nos lleva a la siguiente pregunta: ¿Qué tan buenos serán los vectores de características obtenidos usando wavelets frente a los métodos tradicionales?; el término bueno en este contexto significa que tengan información localizada de los cambios en la señal para distintos niveles de frecuencia, es decir, conocer que tanto aporta en la señal determinado nivel de frecuencia, en determinado tiempo.

Para lograr comprobar si nuestra hipótesis es verdadera, compararemos los vectores de características obtenidos con wavelets, frente a los obtenidos con la Transformada de Fourier, y utilizaremos un reconocedor común en ambos casos.

El mejoramiento del espectro se traduce como mejores tasas de reconocimiento;

se hace también una comparación con unos de los métodos más robustos existentes actualmente para extracción de características , el método MFCC.

La presente tesis está organizada de la siguiente manera, se hace una breve introducción a los sistemas informáticos del lenguaje hablado, su arquitectura, algunos trabajos previos y una breve descripción histórica. Luego se describe el habla así como algunas de sus propiedades físicas para después abordar el procesamiento digital de la señal, donde se muestran algunos conceptos previos aplicables a la extracción de características de la señal de habla; se describe también la técnica de los Coeficientes Cepstrales en Frecuencia Mel. Posteriormente se definen las wavelets. A continuación se muestra el reconocedor basado en Dynamic Time Warping que fué utilizado para medir el desempeño de nuestra técnica basada en las wavelets frente a los Coeficientes Cepstrales en Frecuencia Mel. Finalmente se detalla el procedimiento para extraer características de la señal de habla mediante wavelets, mostrando los resultados obtenidos y las conclusiones de la presente tesis.

La contribución que la presente tesis brinda es la técnica de extracción de características basadas en las wavelets, para la construcción de reconocedores de habla con menor costo computacional que pueden ser aplicados en diversos sistemas informáticos de nuestro medio, por ejemplo en diversas industrias; por este motivo las pruebas se hicieron con voces pertenecientes a personas de nuestra región. Es así como la ciencia es aplicada para crear tecnología, en este caso la aplicación de las

wavelets en sistemas informáticos de reconocimiento automático del habla.

Los Autores.

Parte I
Marco Teórico

Capítulo 1

Sistemas Informáticos del Lenguaje Hablado

Los seres humanos tenemos muchas formas de comunicarnos, pero la más dominante de todas es el habla, pues constituye la más importante manera de intercambiar información. Desde los tiempos primitivos hasta los tiempos modernos, con la aparición de ciertas tecnologías como el teléfono, la radio, la televisión, las películas; es notoria la importancia que tiene la palabra hablada en la psicología humana. Si el habla es la principal forma de comunicación de humano-humano, luego esto motiva a entender que también sea una buena y natural manera de tratar de entablar una comunicación humano-máquina. Las computadoras actuales, nos ofrecen diversas Interfases Gráficas de Usuario, en las cuales por ejemplo los usuarios, pueden hacer click con el mouse o ejecutar un comando con el teclado. Actualmente muchos investigadores, buscan, la posibilidad de que las computadoras hablen, escuchen, entiendan, aprendan e inclusive vean; mostrando resultados prometedores, pero todavía las computadoras actuales carecen de todas estas características propiamente humanas.

El habla puede constituir una interfaz natural entre los humanos y las máquinas por ser esta la manera mas natural como las personas nos comunicamos en la vida diaria, en el trabajo; ya existen muchas aplicaciones en las cuales las tecnologías de los Sistemas Informáticos del Lenguaje Hablado[†] han sido implantadas, como es el caso

[†] del inglés “Spoken Language System”

de celulares, algunas aplicaciones software para usuarios finales, software industrial y de aprendizaje.

Un sistema informático de lenguaje hablado, requiere algunos módulos como: un módulo de Reconocimiento Automático del Habla que convierta las señales acústicas provenientes del aparato fonador del hablante a texto, otro módulo de Entendimiento del Lenguaje Hablado que razone y extraiga el significado de manera coherente del texto generado por el módulo anterior, es decir sea capaz de interpretar el habla; y a la vez sea capaz de reunir el conocimiento necesario en un respectivo dominio para que a través de ciertas reglas, la computadora pueda emitir información y generar una determinada acción y finalmente un módulo de Síntesis del Habla, por el cual la computadora tendrá la capacidad de emitir sonidos, en este caso será el lenguaje hablado de la forma mas natural posible, constituyendo de esta manera, una mínima interfaz con el usuario.

Existen potencialmente dos clases de usuarios quienes se pueden beneficiar con la adopción del habla como una modalidad de control en paralelo con otras modalidades como el mouse, el teclado, pantallas digitales, joystick, etc. La primera clase de usuarios son los usuarios novatos, las funciones que son realmente simples deberían ser accesibles como por ejemplo levantar o disminuir el volumen de los parlantes, bajo un software de control de escritorio, que conceptualmente es una simple operación, en algunas Interfaces de Usuario actuales, requieren el abrir de una a varias ventanas

o menús, manipular sliders, cajas de texto y otros elementos; esto requiere algún conocimiento de convenciones del sistema de interfaces y estructura, para usuarios novatos operaciones sencillas como levantar el volumen o apagar el ordenador, iniciar un determinado programa, etc. deben ser realizadas de la manera mas natural posible. Para usuarios expertos muchas veces las Interfaces Gráficas de Usuario constituyen un obstáculo, muchas veces se requerirá la utilización de las manos del usuario manipulando el teclado o el mouse, y que a la misma vez esté en iteracción con comandos del sistema; por ejemplo un operador de un sistema de diseño gráfico para CAD/CAM, debe especificar un comando en cierto formato de texto, mientras lleva el puntero a cierta área de la pantalla. El habla cumple esas funciones de forma mucho más eficiente que el teclado o los clics del mouse, y se puede de cierta manera diseñar un sistema con interfaces de usuario de forma multimodal, que incluya al habla para capturar aspectos dinámicos del usuario y del estado del sistema; por ejemplo en un entorno donde la parte visual no requiera distracciones y las manos estén ocupadas manejando probablemente un vehículo, el habla podría ser una interfaz entre el hombre y la máquina. Podemos también imaginar un entorno donde dos personas hablen diferente lenguaje, y exista un mecanismo para que las dos se entiendan sin problemas en tiempo real, este es un potencial escenario para una interfaz de lenguaje hablado, que podría tener un módulo de Reconocimiento Automático del Habla, un módulo de síntesis del habla y necesitaríamos un módulo muy sofisticado de entendimiento que

sea multilingüe, si bien todo ello ahora es parte de la ciencia ficción, es una muestra de la importancia de las investigaciones en esta área y sus potenciales aplicaciones.

1.1 Arquitectura

El procesamiento del lenguaje hablado se refiere a la tecnología comprometida con el Reconocimiento Automático del Habla, Síntesis del Habla y Entendimiento del lenguaje Hablado. Un Sistema Informático de Lenguaje Hablado tiene por lo menos uno de los tres módulos mencionados: un módulo de Reconocimiento Automático del habla, que convierte el habla en palabras y las almacena en un ordenador, un módulo de Síntesis del Habla, que hace una conversión texto-habla y origina palabras habladas por el ordenador, y un módulo de Entendimiento del Lenguaje Hablado que convierte las palabras en acciones, asocia conocimiento, y que planifica en el sistema aquellas acciones.

1.1.1 Reconocimiento Automático del Habla

Las ventajas del Reconocimiento Automático del Habla, como vía de dar ordenes a los ordenadores son:

- Permite acceso remoto, al poder acceder a un ordenador usando la red telefónica, que es la red de comunicaciones más extendida.
- Permite la disminución del tamaño de los paneles de control. Piénsese en el panel de un avión, cuantos conmutadores manuales podrían suprimirse si se

utilizara la voz como forma de comunicación con el sistema de control.

- Permite movilidad, ya que la voz se puede enviar a distancia y ser recogida por un micrófono, por oposición a un teclado que no se puede mover de la mesa de trabajo.
- Hace la comunicación más rápida y más agradable para los usuarios, ya que al ser la forma natural de comunicarse no se necesita ninguna habilidad especial.
- Permite tener las manos libres para utilizarlas en alguna otra actividad, a la vez que se van dando ordenes por medio de la voz.

La tecnología del Reconocimiento Automático del Habla es una actividad multidisciplinaria, en la que deben intervenir desde científicos de la computación, ingenieros electrónicos, programadores, hasta psicólogos y lingüistas.

1.1.2 Síntesis del Habla

También llamada conversión Texto-Habla, puede verse como la tarea inversa del Reconocimiento Automático del Habla, y el objetivo es generar sonidos por la computadora que se asemejen a la voz humana, a partir de un texto almacenado en el computador.

Se debe tener en cuenta la entonación de las frases cuando se pronuncian, las pausas entre estas, el orden y duración natural de los sonidos y tratar con un número grande de palabras, convirtiéndose en un problema no trivial.

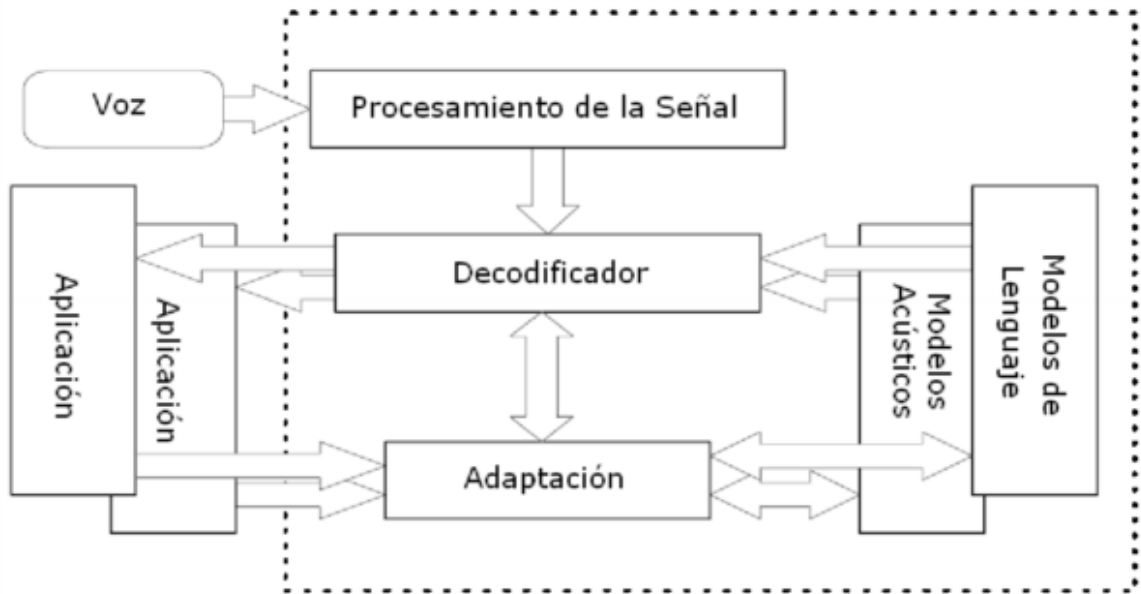


Figura 1.1 Arquitectura básica de un Sistema de Reconocimiento Automático del Habla. Fuente : [Huang and Hon, 2001]

Las investigaciones en el tema datan de la fecha de 1930, y hasta ahora, si bien es cierto se ha podido generar voz humana a través programas de conversión Texto-Habla para computadora, todavía ésta no llega a la calidad de la voz humana, pero existen muchas aplicaciones comerciales.

1.1.3 Entendimiento del Lenguaje Hablado

Un Sistema de Entendimiento del Lenguaje Hablado es necesario para interpretar las expresiones en determinado contexto y para llevar a cabo las acciones apropiadas; el conocimiento léxico, sintáctico y semántico debe ser aplicado de manera que permita una interacción cooperativa entre varios niveles de conocimiento acústico,

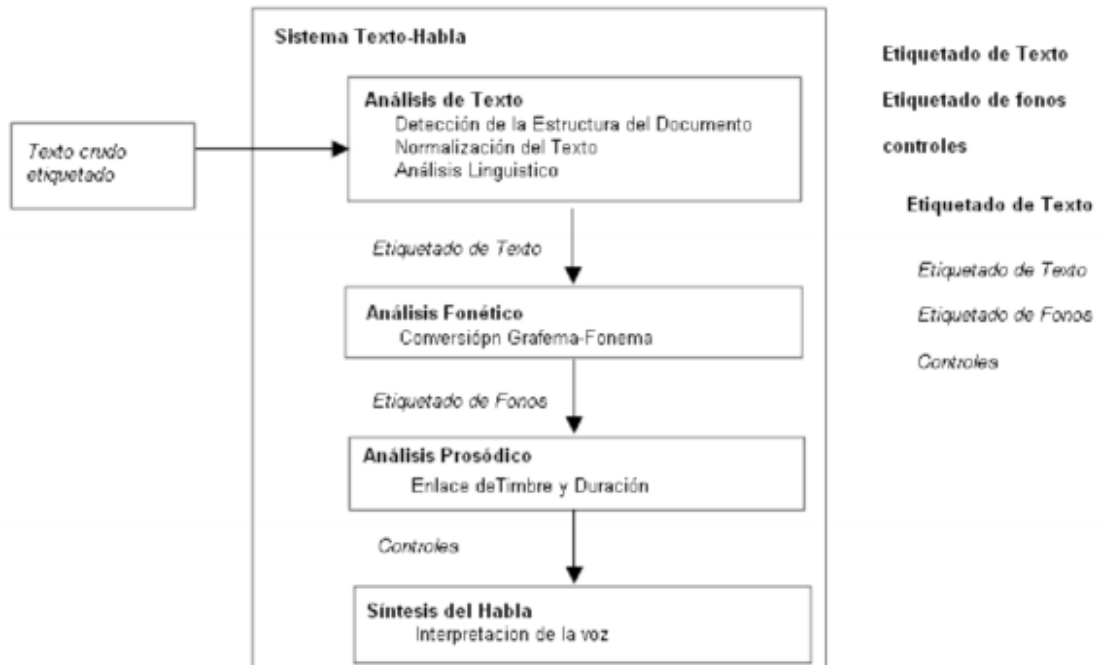


Figura 1.2 Arquitectura básica de un Sistema de Conversión Texto-Habla. Fuente : [Huang and Hon, 2001]

fonético, lingüístico y conocimiento aplicativo para minimizar las incertidumbres.

El conocimiento de las características del vocabulario, los típicos patrones sintácticos y las posibles acciones en un determinado contexto para la interpretación de las expresiones de los usuarios y para la planificación de la actividad del sistema, que es a su vez la parte vital de cualquier Sistema Informático del Lenguaje Hablado.

1.2 Trabajos Previos

Actualmente existe mucha bibliografía en el campo de Reconocimiento Automático del Habla (RAH) y en el Procesamiento Digital de Señales aplicable al RAH, en consecuencia existen muchas técnicas para el RAH, pero a manera común estas técnicas encajan en un modelo de construcción mínimo, que es el siguiente:

La etapa encargada del procesamiento digital de la señal de habla, que vendrá a ser la encargada de capturar la señal analógica de voz, digitalizarla, manipularla mediante diversos algoritmos y brindar un conjunto de coeficientes de características que representen muy bien a una palabra dada, la complejidad de esta etapa está dada por la manera de como se obtiene los mejores coeficientes de características, es decir los coeficientes que mejor representen a una palabra o fonema dado.

Los vectores de características resultantes de la etapa anterior son enviados al reconocedor que será la etapa encargada de hacer una clasificación del patrón entrante, la dificultad existente en esta parte está dada por la complejidad de analizar los vectores de características resultantes, pues se podría decir por ejemplo la palabra "casa" en 1 segundo, y otra persona podría decir la misma palabra en 1.7 segundos y con un timbre de voz, tonalidad y volumen de voz diferente; además la complejidad añadida, si se está tratando de reconocer el habla en una conversación normal es encontrar donde empieza una palabra y donde termina, por ejemplo si se dice "hola como estas" de manera cotidiana (sin pausas), en este contexto donde empieza y

donde termina una palabra no es muy obvio. Estos dos etapas son las más básicas, pues se puede añadir más etapas para mejorar la performance del sistema, como por ejemplo la etapa de modelado de lenguaje, que será la encargada de brindar estructuras léxicas, como palabras de diccionario, y hacer uso de la teoría de lenguajes formales [Huang and Hon, 2001]; otra etapa de modelado acústico, encargada de establecer modelos de producción del habla y modelos de percepción del habla en los humanos; estos modelos son establecidos generalmente por aproximaciones mediante fórmulas matemáticas de los fenómenos físicos involucrados.

Existen diversos trabajos realizados, en la etapa del procesamiento digital de la señal de habla pero que no involucra un análisis mediante Wavelets [Mantha, 1998], existe un trabajo a manera de introducción sobre la posible aplicación de las wavelets en el RAH, [Aboufadel, 2001], los trabajos de Sarikaya, [Ruhi Sarikaya and Hansen, 2001], [Sarikaya and Hansen, 2000], son unos de los más importantes en lo que se refiere a la aplicación de Wavelets en la etapa del procesamiento digital de la señal de habla, entre otros trabajos no menos importantes tenemos: [Beng, 2000], que hace uso de la transformada Wavelet en el reconocimiento de fonemas y el trabajo de, [M. Siafarikas, 2000], que hace uso de la aplicación de Wavelets en el reconocimiento del hablante. En la etapa de reconocimiento uno de los trabajos más importantes es el trabajo de Sakoe and Chiba, [Sakoe and Chiba, 1978] que muestran un algoritmo optimizado para el reconocimiento de palabras haciendo uso de la programación

dinámica.

1.3 Historia

Años 1940's y 1950's , procesamiento digital de la señal muy simple, se detectaba la energía en varios bandos de frecuencia, se introdujeron muchas ideas que son usados en los sistemas actuales de Reconocimiento Automático del Habla como el entrenamiento estadístico y el modelado del lenguaje, también los reconocedores tenían un pequeño vocabulario como dígitos, vocales, etc., generalmente los sistemas no se probaban con muchos hablantes (aproximadamente 10).

1970's , un ambicioso proyecto para construir un sistema informático que entienda y procese el habla fue iniciado por DARPA [Jelinek, 1998], la meta era integrar conocimiento acerca del habla, lingüística e inteligencia artificial para desarrollar un Sistema Informático del Lenguaje Hablado.

Se desarrolló un sistema informático llamado Harpy que integraba todas las fuentes de conocimiento en redes de estado finito, que eran entrenadas estadísticamente.

En estos años crecen notoriamente métodos que hacen uso de la probabilidad y se entiende por Reconocimiento Automático del Habla como buscar la palabra mas probable en una señal de audio, dada alguna información de su distribución de probabilidad.

1980's , modernos reconocedores son lanzados al mercado, los algoritmos desarrollados en este tiempo son usados todavía en nuestros días como: modelos n-grams, mixturas gaussianas, modelos ocultos de Markov, decodificador Viterbi, etc.

En 1984 es construido el primer sistema de dictado en tiempo real, por IBM.

1990's , con el aumento del poder de cómputo y capacidad de memoria de las computadoras existen avances en algoritmos de adaptación, entrenamiento discriminativo; se desarrollan a la vez sistemas informáticos que reconocían palabras independientemente quien fuera el hablante, algunas aplicaciones fueron implantadas en empresas telefónicas.

1995 , Dragon IBM lanza su producto que reconocía palabras aisladas dependiente del hablante, era un sistema que permitía un dictado de un gran número de palabras.

1997 Dragon IBM lanza su sistema para dictado continuo.

Actualidad , se espera que con el poder de cómputo actual de las máquinas y las diversas investigaciones en el tema, se logre aumentar el desempeño. Aplicaciones actuales van desde software para celulares, robótica hasta interfaces de voz para personas discapacitadas; desafortunadamente las comparaciones de error entre humanos y máquinas todavía dan un amplio margen de diferencia, se espera lograr que éste margen de diferencia disminuya en los próximos años.

Tarea	Vocabulario	Humanos	Máquinas
Dígitos conectados	10	0.009%	0.72%
Letras del alfabeto	26	1%	5%
Habla espontánea por teléfono	2000	3.8%	36.7%

Tabla 1.1 Error humano es aproximadamente 5 veces menor que el error de las máquinas. Fuente : [Huang and Hon, 2001]

Capítulo 2

El Habla, Producción y Percepción

El campo de estudio del Reconocimiento Automático del habla por parte de un ordenador, ha sido y es abordado como un trabajo interdisciplinario entre científicos de la computación y profesionales de otras áreas en la búsqueda de algoritmos para reconocimiento de patrones, gramáticas regulares, analizadores léxicos sintácticos, etc.; estos algoritmos deben ser óptimos por el elevado número de valores a tratar, es decir se debe buscar el mejor algoritmo que resuelva una situación determinada en el menor tiempo posible y que a la vez consuma pocos recursos, es por eso que el diseño y análisis de algoritmos se vuelve un pilar fundamental en este campo de investigación, convirtiéndose un trabajo interdisciplinario.

A continuación presentamos parte de la teoría fundamental para el desarrollo de esta tesis.

2.1 El Habla como Sonido

El sonido es una onda que transporta su energía en forma paralela a su movimiento, está formada por compresiones y rarefacciones de moléculas de aire; cuando producimos un sonido, independientemente del tipo que sea, provocamos una perturbación en las moléculas de aire, dicha perturbación es captada por nuestros oídos, que lo interpretan como un sonido en particular, en consecuencia el sonido tiene una doble

interpretación, por una parte para los físicos será una perturbación en las moléculas de aire, independientemente de la sensación que producen; pero para los fisiólogos será sonido todo aquello que resulta audible por el sistema auditivo humano, entonces se tiene un punto de vista fisiológico y físico del sonido; en conclusión para esta tesis diremos que sonido es todo aquello que produce una perturbación física en las moléculas de aire causando compresión y rarefacción, es decir un movimiento de vaivén en las moléculas medidas en su punto de origen, movimiento que es transmitido en forma de cadena a todas las demás moléculas de aire y que causan una sensación auditiva en un determinado receptor.

El habla es el sonido que emite el aparato fonador humano, pero no es cualquier sonido, si no que contiene además información perteneciente a un determinado lenguaje o idioma.

El sonido al tener un movimiento ondulatorio en las moléculas de aire originará un cambio de presión y una transmisión de energía en forma paralela al movimiento de éstas. La velocidad del sonido a una temperatura de 0.6 grados centígrados es de 331.5 m/s.

La cantidad de trabajo requerida para generar la energía está en función del grado de desplazamiento de las moléculas de aire de su posición de reposo, este grado de desplazamiento es llamado amplitud de un sonido, el amplio rango de los valores de amplitud hace que ésta se mida en una escala logarítmica (decibeles), la escala decibel

es la comparación entre dos sonidos:

$$10 \log_{10} \frac{P_1}{P_2}. \quad (2.1)$$

Donde P_1 y P_2 son dos niveles de presión.

La medida absoluta del nivel de presión de un sonido SPL , es una medida absoluta de la presión P del sonido en decibeles:

$$SPL(dB) = 20 \log_{10} \frac{P}{P_0}. \quad (2.2)$$

Donde 0 $dB SPL$ corresponde al umbral de audición del oído humano y que corresponde al valor de $P_0 = 0.0002 \mu Bar$ para un tono de 1 Khz , por ejemplo un nivel de conversación normal a 1.5 *metros* es de 3 $dB SPL$, y el sonido producido por un avión es de aproximadamente de 120 $dB SPL$.

2.2 Producción del Habla

El aire que proviene de los pulmones hace vibrar las cuerdas vocales una y otra vez, durante el sonido del habla, luego se dirá que el sonido es sonoro si las cuerdas oscilan una y otra vez, si los pliegues de las cuerdas vocales se encuentran demasiado flojos o demasiado tensos para vibrar periódicamente se dice que los sonidos son no sonoros, el lugar donde las cuerdas vocales vibran se llama glotis.

Luego el aire pasa por el velo del paladar que actúa como una válvula, acá se producen sonidos como la "m" y "n". El paladar duro actúa junto con la lengua para

formar sonidos de diversas consonantes, la lengua en diversas posiciones puede dar lugar a consonantes y vocales, los dientes también son juntados para la formación de ciertas consonantes, finalmente la forma de disposición de los labios va a dar lugar a las vocales, si los labios son cerrados completamente darán lugar a ciertas consonantes como la "p", "b", "m".

Los sonidos sonoros como las vocales tienen mayor energía que los no sonoros, los diversos timbres son formados por la forma en que la lengua y los labios son dispuestos, también la resonancia formada por la cavidad oral. Las cuerdas vocales vibran a razón de 30 Hz en un hombre adulto hasta 300Hz en un niño o una mujer

La tasa de frecuencia de las cuerdas vocales (abierto-cerrado) en la laringe, va a llamarse frecuencia fundamental y diversos armónicos de alta frecuencia van a acompañar a esta frecuencia debido a las resonancias producidas por la cavidad oral. La frecuencia fundamental también contribuye más que cualquier otro factor para la percepción del tono (la caída y afloramiento semi-musical de tonos de voz) en el discurso.

2.3 Percepción del Habla

Existen dos componentes en el sistema de percepción, los órganos periféricos auditivos (oídos) y el sistema del nervio auditivo (cerebro).

En el oído es donde se procesa la información acústica y luego la información resultante es enviada al cerebro a través del nervio auditivo para su procesamiento.

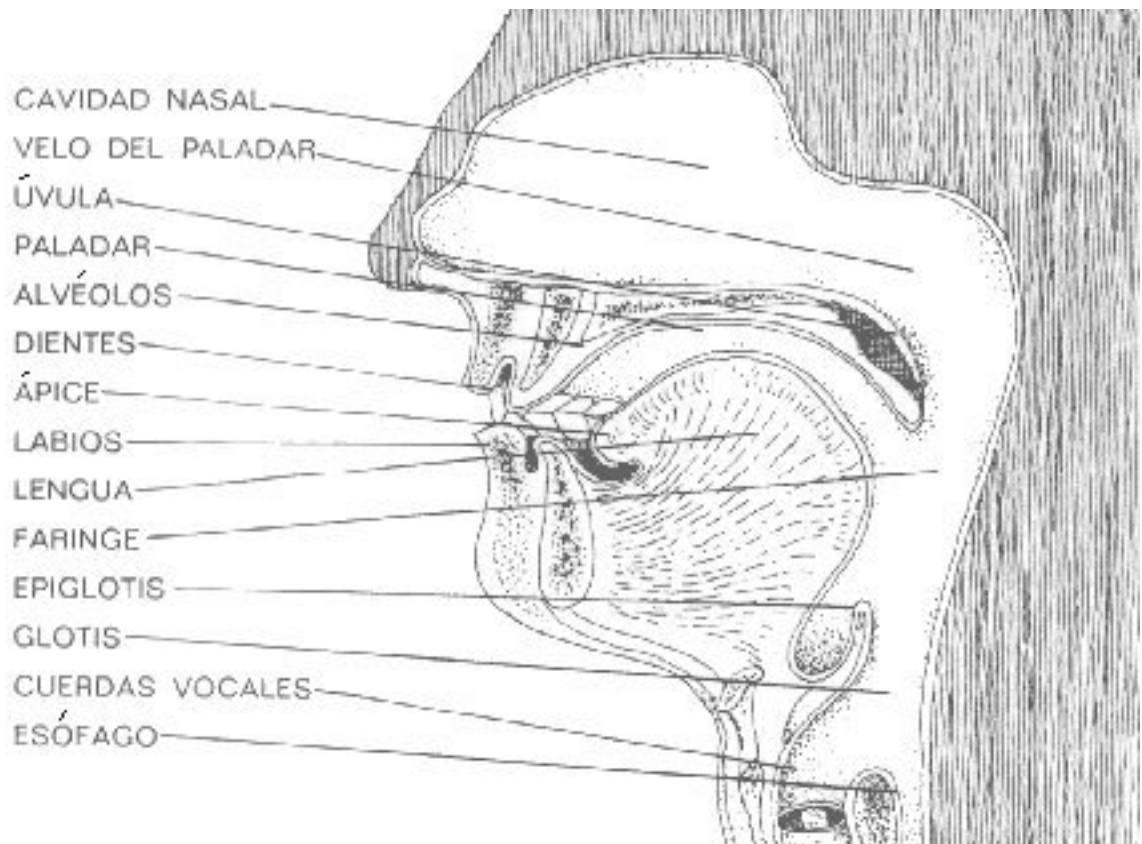


Figura 2.1 Diagrama del aparato fonador humano.

La manera en que los humanos captamos la información acústica por medio de los oídos y como el cerebro procesa dicha información, inspira la construcción de modelos informáticos, el objetivo esta tesis es construir un modelo basado en la percepción del habla utilizando funciones matemáticas llamadas wavelets, pues en el oído interno se produce una descomposición en frecuencias de la señal de una manera muy parecida al tratamiento con wavelets de una señal.

2.3.1 Fisiología del Oído

El oído está dividido en tres partes: oído externo, oído medio y oído interno; el oído externo es el encargado de percibir y canalizar las señales acústicas al interior del oído, tiene una función receptora de señales y hace una normalización del sonido, es decir los sonidos con mucha amplitud los reduce y los sonidos con muy baja amplitud los incrementa, esto ayuda a que el oído escuche sonidos muy bajos y que ablande de alguna manera los sonidos fuertes [Bernal, 2000]; en el oído externo existe el pabellón auricular que es el encargado de percibir los sonidos y dirigirlos hacia el conducto auditivo externo que tiene un tamaño de 25 a 30 mm, parte de su función es la de proteger al oído externo, tiene además una función de resonancia, que es lo que permite en cierto modo el incremento de las señales débiles, también es el encargado de producir cerumen para la lubricación y protección del oído.

El oído medio tiene la función de incrementar la percepción sonora, tiene una interface aire-liquido y es aquí donde se encuentra el tímpano que también permite una amplificación del sonido por su gran tamaño; cuando las presiones del aire varían, hacen que el tímpano vibre, esta vibración es comunicada a unos huesillos llamados martillo, estribo y yunque, que transmiten a su vez las vibraciones al oído interno; gracias al tímpano, el oído humano gana unos 25 a 30 Db de presión, también atenúa presiones muy fuertes, pues activa un mecanismo de freno que protege las células ciliadas del oído interno.

En el oído interno se encuentra un conducto en forma de caracol llamado cóclea, éste contiene un líquido llamado perilinfa, que es estimulado por el movimiento proveniente del tímpano hacia los huesillos, este estímulo es llamado onda viajera y tiene una vibración dependiendo de la frecuencia de estímulo, es decir, las frecuencias altas estimulan con mayor intensidad la parte basal de la cóclea, que es la parte más amplia y contiene mayor cantidad de células ciliadas, aproximadamente unas 12000, las frecuencias más graves estimulan mejor la parte más interna de la cóclea que tiene menos células ciliadas, lo que ocurre allí, nos hace pensar que internamente el oído hace un análisis de la señal de entrada muy parecido al análisis con wavelets, teniendo más detalles para las altas frecuencias que para las bajas frecuencias.

Finalmente ocurre una conversión de energía mecánica a eléctrica por medio de las células ciliadas, con el correspondiente envío de la información a las redes neuronales del cerebro.

Finalmente se puede decir que los atributos físicos y los atributos perceptuales se pueden relacionar de la siguiente manera:

La intensidad estará relacionada con la fuerza de sonido percibida, la frecuencia fundamental de una señal de voz estará en relación con el tono percibido de una persona al momento de hablar y la forma espectral de una señal tendrá relación con el timbre percibido.

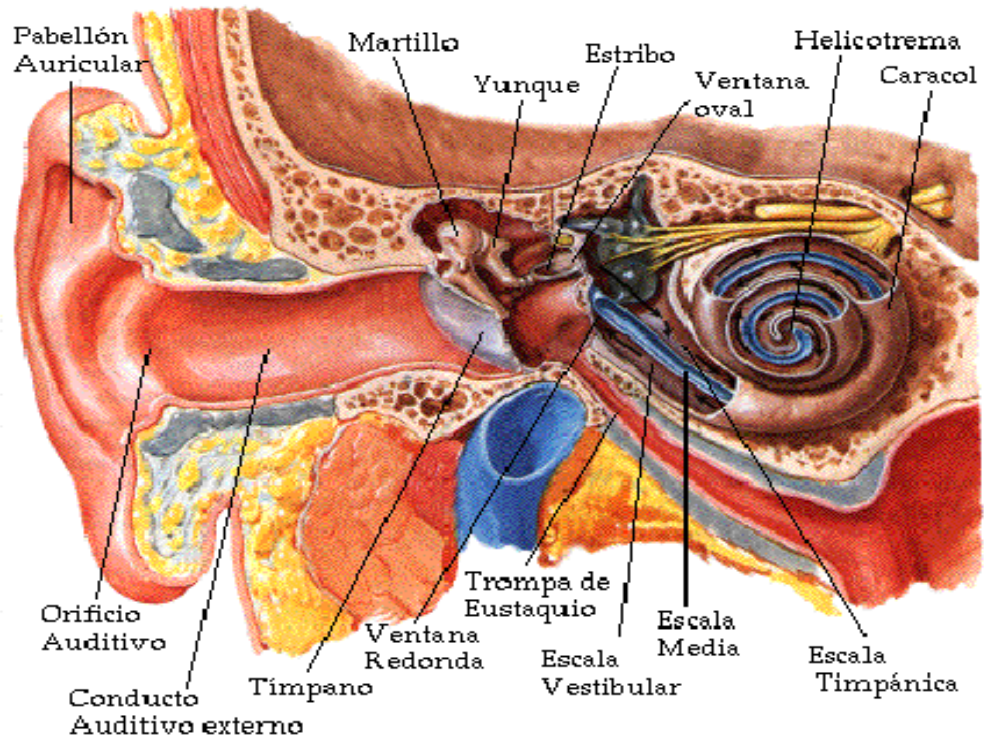


Figura 2.2 Diagrama del sistema auditivo humano. Fuente : Microsoft Encarta 2006.

Cantidad Física	Cantidad Perceptual
Intensidad	Fuerza de Voz
Frecuencia Fundamental	Tono
Forma Espectral	Timbre

Tabla 2.1 Relación entre atributos físicos y perceptuales del sonido. Fuente : [Huang and Hon, 2001]

Capítulo 3

Procesamiento Digital de la Señal

El Procesamiento Digital de Señales juega un rol importante en asegurar que la información proveniente de la señal de habla pueda ser fácilmente extraída por la computadora. Generalmente este trabajo con la señal se hace en el dominio de la Frecuencia, la representación de las señales en este dominio se hace debido a que la estructura de un fonema es generalmente única, tomando en cuenta que si una señal de habla es llevada al dominio de la frecuencia entonces se puede obtener su energía en cada nivel de frecuencia.

El procesamiento digital de la Señal nos permite realizar la extracción de características mediante diversos algoritmos, que finalmente hacen una reducción significativa del tamaño de los datos de entrada.

3.1 Señales Digitales

El Habla puede entenderse como una señal que es representada matemáticamente como una función continua de variable t que representa al tiempo, luego para poder procesar esta señal por medio de una computadora es necesario tener la forma discreta de la señal continua, una señal digital es aquella resultante de hacer un muestreo con Periodo T a la señal continua, luego diremos que las señales digitales también

conocidas como señales discretas en el tiempo, son señales de la forma:

$$x[n] = x_0(nT). \quad (3.1)$$

y se puede definir la frecuencia de muestreo como sigue:

$$F_s = \frac{1}{T}. \quad (3.2)$$

Una de las señales más importantes es el senoide u onda seno:

$$x_0[n] = A_0 \cos(\omega_0 n + \phi_0). \quad (3.3)$$

esta señal es importante pues las señales de habla pueden ser descompuestas como una suma de sinusoides, donde A es la amplitud del senoide, ω_0 es la frecuencia angular y ϕ_0 es la fase; todos estos ángulos expresados en radianes. Operar con identidades trigonométricas puede ser algo tedioso, en consecuencia usaremos una notación basada en el uso de números complejos.

Un número complejo se representa de la forma:

$$z = x + jy. \quad (3.4)$$

donde x es la parte real, y es la parte imaginaria y $j = \sqrt{-1}$; también un número complejo tiene una representación polar de la forma:

$$z = Ae^{j\phi}. \quad (3.5)$$

donde $x = A \cos \phi$ y $y = A \sin \phi$, luego usando la relación de Euler se tiene:

$$e^{j\phi} = \cos \phi + j \sin \phi \quad (3.6)$$

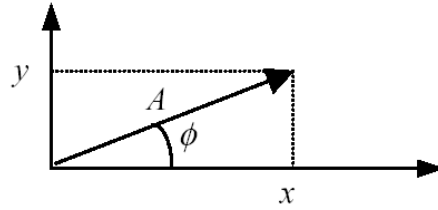


Figura 3.1 Representación de un número complejo en el plano cartesiano.

Un número complejo z en su forma polar $z = Ae^{j\phi}$, donde A es la amplitud y ϕ es la fase, luego si se usan números complejos el senoide puede ser representado como sigue:

$$x_0[n] = A_0 \cos(\omega_0 n + \phi_0) = \text{Re}\{A_0 e^{j(\omega_0 n + \phi_0)}\} \quad (3.7)$$

3.2 Sistemas Digitales

Se dice que un sistema es digital si:

$$y[n] = T\{x[n]\} \quad (3.8)$$

Serán lineales si:

$$T\{a_1 x_1[n] + a_2 x_2[n]\} = a_1 T\{x_1[n]\} + a_2 T\{x_2[n]\} \quad (3.9)$$

Serán invariantes en el tiempo si:

$$y[n - n_0] = T\{x[n - n_0]\} \quad (3.10)$$

Luego se puede definir la operación convolución y su propiedad conmutativa como sigue:

$$y[n] = \sum_{k=-\infty}^{\infty} x[k]h[n - k] = \sum_{k=-\infty}^{\infty} x[n - k]h[k] \quad (3.11)$$

puede ser escrita como:

$$y[n] = x[n] * h[n] \quad (3.12)$$

3.3 Transformada de Fourier

La Transformada de Fourier es una herramienta de análisis que transforma una señal representada en el dominio del tiempo hacia el dominio de la frecuencia, sin alterar su información.

La Transformada de Fourier está dada por:

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t} \quad (3.13)$$

y su inversa llamada Transformada Inversa de Fourier está dada por:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega)e^{j\omega t} \quad (3.14)$$

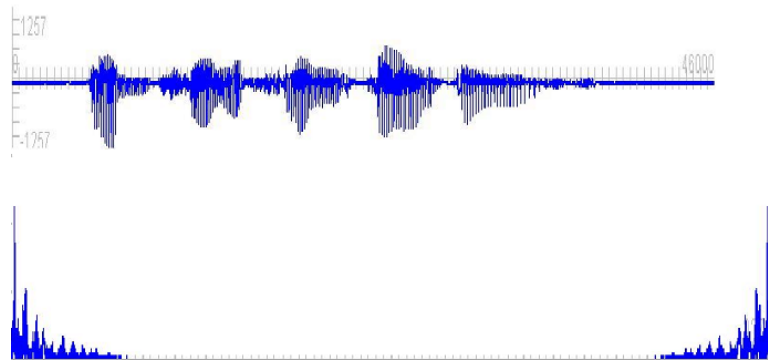


Figura 3.2 En la parte superior se tiene la señal de habla en el dominio del tiempo. En la parte inferior se tiene la señal de habla en el dominio de la frecuencia.

3.3.1 Transformada Discreta de Fourier

Transforman una señal discreta en el tiempo en una señal discreta en la frecuencia, la Transformada Discreta de Fourier de una señal esta definida por:

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-\frac{j2\pi nk}{N}} = \sum_{n=0}^{N-1} x[n]W_N^{kn}, \quad k = 0, 1, \dots, N-1 \quad (3.15)$$

donde:

$$W_N = e^{-\frac{j2\pi}{N}} \quad (3.16)$$

El factor W es llamado también *factor mariposa*, el cual es una función de N términos de frecuencia, con argumento nk

La Transformada Inversa Discreta de Fourier esta dada por:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k]e^{\frac{j2\pi nk}{N}} = \sum_{k=0}^{N-1} X[k]W_N^{-kn} \quad n = 0, 1, \dots, N-1 \quad (3.17)$$

3.3.2 Complejidad Computacional de la Transformada Discreta de Fourier

La Transformada Discreta de Fourier tiene la siguiente ecuación de recurrencia

$$T[n] = \begin{cases} T(n-1) + N & \text{si } 0 < n \leq N \\ 0 & n = 0 \end{cases} \quad (3.18)$$

resolviendo la ecuación de recurrencia

$$T(n) = T(n-1) + N$$

$$T(n-1) = (T(n-2) + N) + N$$

$$\dots = \dots$$

$$T(n-n+1) = T(n-n) + N + N + \dots + \dots + N$$

$$T(n) = \underbrace{NxN}$$

$$T(n) = N^2$$

(3.19)

se tiene una complejidad $O(n^2)$

3.3.3 Transformada Rápida de Fourier

Utilizar la Transformada Discreta de Fourier tiene un costo computacional elevado, cuando se trabaja con muchos datos, para ello un algoritmo propuesto por [Cooley and Tukey, 1965] hizo que el cálculo de la Transformada Discreta se realizara en forma mas eficiente, este es el algoritmo de la Transformada Rápida de Fourier, el aporte está en el hecho de que la Transformada Discreta de Fourier necesita N^2 operaciones mientras que la Transformada Rápida de Fourier solo necesita $N \log_2 N$ operaciones, donde N es el número de muestras, por ejemplo si $N = 1024$, la Transformada Discreta de Fourier necesitaría $1024^2 = 1,048,576$ operaciones entre multiplicaciones y sumas como mínimo, mientras que la Transformada Rápida de Fourier solo necesitaría $1024 \log_2 1024 = 10,240$ operaciones entre multiplicaciones y sumas como mínimo; la ventaja del segundo algoritmo es evidente. Existen muchos algoritmos que implementan la Transformada Rápida de Fourier y estos se basan en el paradigma *divide y conquista* [Cooley and Tukey, 1965], los cuales primero dividen el problema en dos o más subproblemas de pequeño tamaño, solucionan el mismo

subproblema recursivamente por el mismo algoritmo, aplicando condiciones de limite para terminar la recursion cuando el tamaño de los subproblemas son suficientemente pequeños y obtienen la solución al problema original, combinando las soluciones de los subproblemas.

Existen muchos algoritmos Rápidos de Fourier y su aplicación depende de criterios, como el tamaño de datos a tratar, si el algoritmo va a ser implementado en computadoras con un procesador o en computadoras con varios procesadores, etc.; para el presente trabajo, hemos optado en utilizar el algoritmo *Radix - 2 FFT*, el cual tiene dos variantes habitualmente usadas [Chu and George, 2000], los cuales son algoritmos secuenciales es decir se van a implementar en computadoras con un solo procesador, estos algoritmos son: algoritmo *Radix - 2 con Decimacion en el Tiempo* y algoritmo *Radix - 2 con Decimacion en la Frecuencia*, solamente varían en el modo en como los dos subproblemas son definidos, no existiendo ninguna variación en complejidad computacional.

3.3.4 Algoritmo Radix-2 con Decimación en Frecuencia y reordenamiento en la salida de bits mezclados

Los primeros algoritmos rápidos fueron propuestos en forma independiente por [Cooley and Tukey, 1965] en su forma de decimación en el tiempo y en su forma de decimación en frecuencia por [Gentleman and Sande, 1966]. A continuación se describe el algoritmo Radix-2 con decimación en frecuencia y reordenamiento en la

salida de bits.

Tenemos la Transformada Discreta de Fourier definida de la siguiente manera:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N} = \sum_{n=0}^{N-1} x[n] W_N^{kn}, \quad k = 0, 1, \dots, N-1 \quad (3.20)$$

donde:

$$W_N = e^{-j2\pi/N} \quad (3.21)$$

donde podemos deducir lo siguiente:

$$W_N^{N/2} = -1$$

$$W_{N/2} = W_N^2$$

$$W_N^N = 1$$

donde N es necesariamente un valor que sea el resultado de una potencia de dos, para que el algoritmo pueda aplicar eficientemente el principio de *divide y conquista*.

El algoritmo radix-2 con decimación en la frecuencia, es obtenido por definir dos subproblemas de la forma:

$$\{X(2k) | k = 0, \dots, \frac{N}{2} - 1\} \quad (3.22)$$

que son las series de frecuencias de salidas diezmadas en índices pares y:

$$\{X(2k+1) | k = 0, \dots, \frac{N}{2} - 1\} \quad (3.23)$$

las series de frecuencias de salidas diezmadas en índices impares; para lograr esto la ecuación (3.21) puede ser escrita como[†]

[†] usaremos X_k por $X(K)$ y X_n por $x(n)$

$$X_k = \sum_{n=0}^{\frac{N}{2}-1} x_n W_N^{kn} + \sum_{n=\frac{N}{2}}^{N-1} x_n W_N^{kn} \quad (3.24)$$

$$X_k = \sum_{n=0}^{\frac{N}{2}-1} x_n W_N^{kn} + \sum_{n=0}^{\frac{N}{2}-1} x_{n+\frac{N}{2}} W_N^{k(n+\frac{N}{2})} \quad (3.25)$$

$$X_k = \sum_{n=0}^{\frac{N}{2}-1} (x_n + x_{n+\frac{N}{2}} W_N^{k\frac{N}{2}}) W_N^{kn}, \quad k = 0, 1, \dots, N-1 \quad (3.26)$$

escogiendo los indices pares tenemos:

$$X_{2k} = \sum_{n=0}^{\frac{N}{2}-1} (x_n + x_{n+\frac{N}{2}} W_N^{kN}) W_N^{2kn} \quad (3.27)$$

$$X_{2k} = \sum_{n=0}^{\frac{N}{2}-1} (x_n + x_{n+\frac{N}{2}}) W_N^{kn}, \quad k = 0, 1, \dots, \frac{N}{2} - 1 \quad (3.28)$$

definiendo $Y_k = X_{2k}$ y $y_n = x_n + x_{n+\frac{N}{2}}$, $k = 0, 1, \dots, \frac{N}{2} - 1$, tenemos la primera mitad del problema

$$Y_k = \sum_{n=0}^{\frac{N}{2}-1} y_n W_N^{kn}, \quad k = 0, 1, \dots, \frac{N}{2} - 1 \quad (3.29)$$

similarmente para los indices impares tenemos:

$$X_{2k+1} = \sum_{n=0}^{\frac{N}{2}-1} (x_n + x_{n+\frac{N}{2}} W_N^{(2k+1)\frac{N}{2}}) W_N^{(2k+1)n} \quad (3.30)$$

$$X_{2k+1} = \sum_{n=0}^{\frac{N}{2}-1} ((x_n - x_{n+\frac{N}{2}}) W_N^n) W_N^{kn}, \quad k = 0, 1, \dots, \frac{N}{2} - 1 \quad (3.31)$$

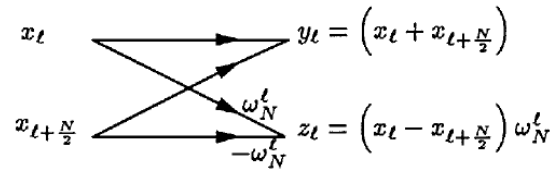


Figura 3.3 La mariposa Gentleman-Sande.

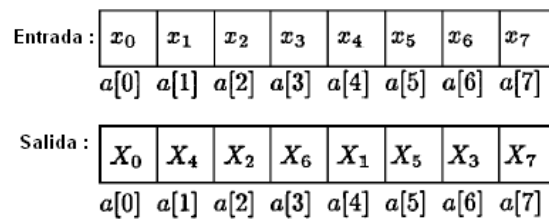


Figura 3.4 La entrada x en el array a es sobrescrita por la salida mezclada X .

definiendo $Z_k = X_{2k+1}$ y $z_n = (x_n - x_{n+\frac{N}{2}})W_N^n$, $k = 0, 1, \dots, \frac{N}{2} - 1$, obtenemos la segunda mitad del problema

$$Z_k = \sum_{n=0}^{\frac{N}{2}-1} z_n W_{\frac{N}{2}}^{kn}, \quad k = 0, 1, \dots, \frac{N}{2} - 1 \quad (3.32)$$

El cálculo de y_n y de z_n en el paso de la subdivision es definida en la literatura como: la mariposa Gentleman-Sande.

La salida de este procedimiento obtiene los valores de frecuencia pero en un orden mezclado, es decir si se tiene los valores de entrada x almacenados en un array a , la salida cuando el algoritmo Radix-2 con Decimación en Frecuencia sea computado, obtendremos los valores X en el array a pero de una manera mezclada.

El reordenamiento de estos valores se hace teniendo en cuenta lo siguiente: las posiciones ordenadas de los valores de salida serán obtenidos al hacer un operación de *bits reverso* a las posiciones del vector de salida.

Si la notación binaria de un número A es la siguiente:

$$A = abcd\dots yz \quad (3.33)$$

donde $abcde\dots yz$ son valores que corresponden a los números 0 o 1, su *bits reverso* sera:

$$A = zyxw\dots a \quad (3.34)$$

que corresponden a intercambiar la última posición de su representación binaria por la primera, la penúltima por la segunda, la antepenúltima por la tercera y así sucesivamente.

3.3.5 Complejidad Computacional de la Transformada Rápida de Fourier

La Transformada Rápida de Fourier tiene la siguiente ecuación de recurrencia:

$$T[n] = \begin{cases} 2T(n/2) + Cn & \text{si } 2^n \geq 2 \\ 0 & n = 1 \end{cases} \quad (3.35)$$

resolviendo la ecuación de recurrencia

$$\begin{aligned} T(n) &= 2T(n/2) + Cn \\ T(n/2) &= 4T(n/4) + 2C\frac{n}{2} + Cn \end{aligned}$$

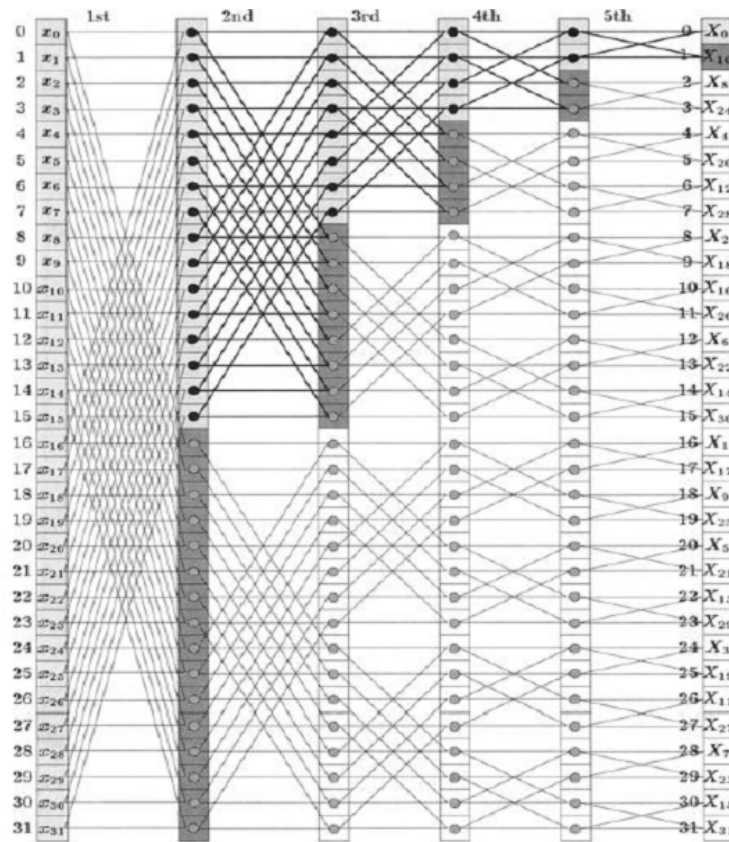


Figura 3.5 Algoritmo Radix 2 . Fuente: [Chu and George, 2000]

... = ...

$$2^a = n$$

$$a = \log_2 n$$

... = ...

$$2^{a-1}T(n/2^{a-1}) = 2^aT(n/2^a) + \underbrace{2^{a-1}C \frac{n}{2^{a-1}} + 2^{a-2}C \frac{n}{2^{a-2}} + \dots + \dots + Cn}_{}$$

$$2^{a-1}T(n/2^{a-1}) = n + a(cn)$$

$$T(n) = n + cn \log_2 n \quad (3.36)$$

La Transformada Rápida de Fourier tiene una complejidad de $O(n \log n)$

3.4 Función Ventana

Las funciones ventana son señales concentradas en un lapso de tiempo, para lograr así enfocar un mayor análisis en cierta región en particular y tratar de evitar las discontinuidades al principio y al final de los bloques analizados, existen muchas funciones ventana como las triangulares, Kaiser, Barlett entre muchas otras, las que son mayormente usadas en los sistemas digitales de procesamiento del habla son la ventana rectangular, Hanning y Hamming.

La utilización de una ventana cambia el espectro en frecuencia en la señal.

3.4.1 Ventana Rectangular

La ventana rectangular es definida como:

$$h_n = u[n] - u[n - N] \quad (3.37)$$

donde la función u es la función paso unitario definida como:

$$u[n] = \begin{cases} 1 & \text{si } n \geq 0 \\ 0 & \text{si } n < 0 \end{cases} \quad (3.38)$$

y N es el tamaño de la ventana.

$$w[n] = 1, \quad 0 \leq n \leq N - 1$$

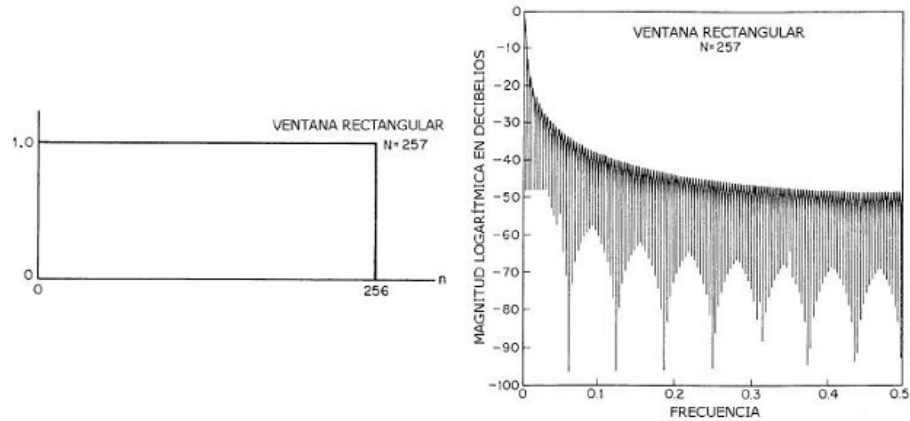


Figura 3.6 Ventana Rectangular en el dominio del tiempo y de la frecuencia.

3.4.2 Ventana Generalizada Hamming

La forma generalizada de la ventana Hamming es definida como:

$$h[n] = \begin{cases} (1 - \alpha) - \alpha \cos\left(\frac{2\pi n}{N}\right) & \text{si } 0 \leq n < N \\ 0 & \text{otra manera} \end{cases} \quad (3.39)$$

Cuando $\alpha = 0.5$ la función ventana es conocida como ventana Hanning y cuando $\alpha = 0.46$ la función es llamada ventana Hamming.

Las ventanas rectangulares raramente son usadas para analizar segmentos de habla, pues si bien tienen alta resolución en el tiempo, producen efectos no deseables en las frecuencias obtenidas, mayormente son usadas las ventanas tipo Hamming o Hanning que producen un menor derramamiento espectral que es un efecto asociado

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1$$

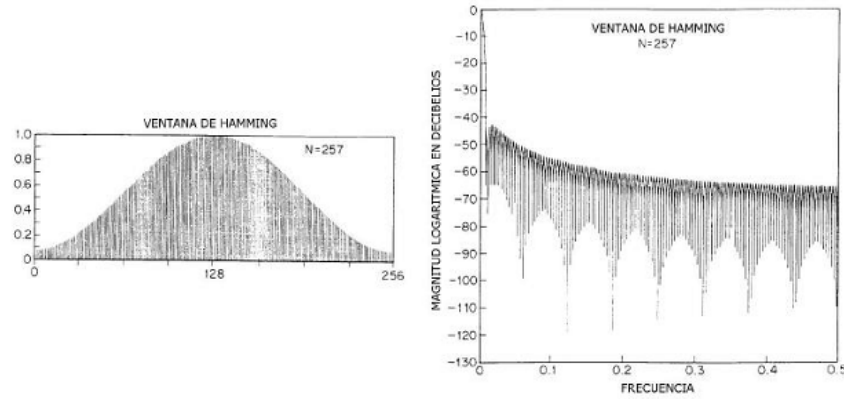


Figura 3.7 Ventana Hamming en el dominio del tiempo y de la frecuencia.

a las ventanas rectangulares.

3.5 Representación de la Señal de Habla

En esta sección se mostrará como obtener una representación de la señal de habla, es decir como obtener valores característicos.

3.5.1 Transformada Corta de Fourier

Descomponer la señal en una serie de segmentos y analizar los segmentos independientemente. Dada una señal $x[n]$ se define una señal corta en el tiempo $x_m[n]$ de un segmento m como sigue:

$$x_m[n] = x[n]w_m[n] \quad (3.40)$$

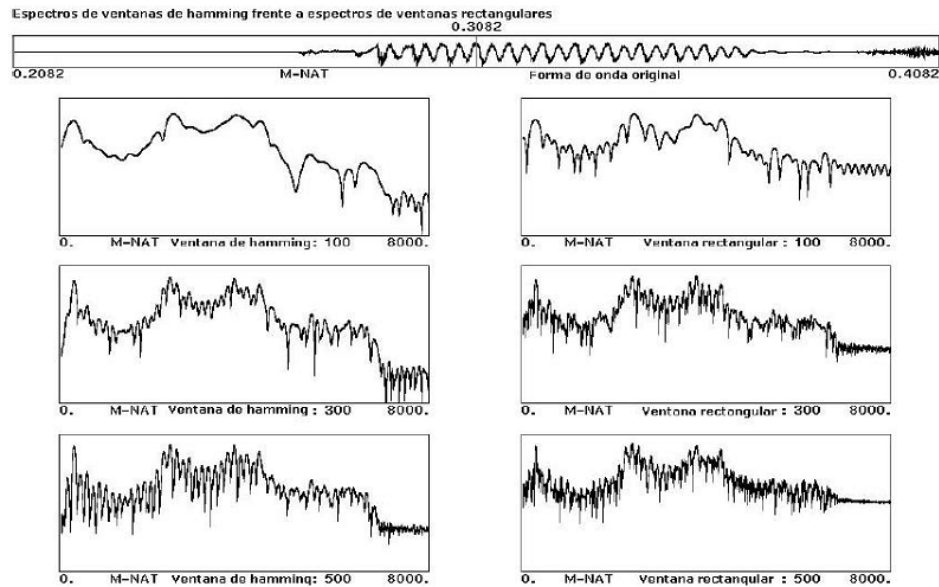


Figura 3.8 Comparación en el dominio de la frecuencia entre la Ventana Rectangular y la Ventana Hamming.

que es el producto de de $x[n]$ por una función ventana $w_m[n]$, luego podremos hacer que la función tenga valores constantes para todos los segmentos:

$$w_m[n] = w[m - n] \quad (3.41)$$

Usualmente se usa un tamaño de ventana de $20ms$ a $30ms$.

Finalmente se tiene que la Transformada Corta de Fourier para un segmento m está definida como:

$$X_m(e^{j\omega}) = \sum \chi_m[n] e^{-j\omega n} = \sum w[m - n] \chi[n] e^{-j\omega n} \quad (3.42)$$

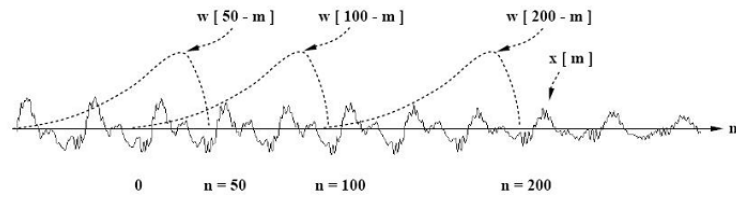


Figura 3.9 Transformada Corta de Fourier.

3.5.2 Transformada Discreta del Coseno

La Transformada Discreta del Coseno (DTC) es ampliamente usada en el procesamiento del habla, ésta tiene varias definiciones; la DTC-II $C[k]$ de una señal real $x[n]$ esta definida:

$$c(m) = \sum_{n=0}^{N-1} x[n] \cos(\pi k (\frac{n + \frac{1}{2}}{N})), \quad 0 \leq k < N \quad (3.43)$$

y su inversa definida por:

$$x[n] = \frac{1}{N} \{ C[0] + 2 \sum_{k=1}^{N-1} C[k] \cos(\pi k (\frac{n + \frac{1}{2}}{N})) \} \quad (3.44)$$

La DTC-II puede ser derivada de la Transformada Discreta de Fourier, la DTC-II es más usada, pues tiene su energía compacta esto quiere decir que concentra su energía en los componentes bajos de frecuencia.

3.5.3 Procesamiento Cepstral

La fuente de excitación $e[n]$ representa a la frecuencia fundamental que se produce en las cuerdas vocales y el filtro $h[n]$ representa las resonancias del tracto vocal dadas por los los labios, faringe, dientes, paladar, etc., que cambian sobre el tiempo.

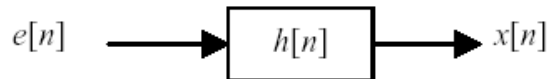


Figura 3.10 Modelo básico fuente-filtro para señales de habla.

La idea es convertir la convolución :

$$x[n] = e[n] * h[n] \quad (3.45)$$

en una suma:

$$\hat{x}[n] = \hat{e}[n] + \hat{h}[n] \quad (3.46)$$

Se define Cepstrum como una transformación que nos va a permitir separar la fuente de excitación del filtro (glotis), se asumirá un valor N para el cual el cepstrum del filtro (glotis) $\hat{h}[n] \approx 0$ para $n \geq N$ y el el cepstrum de la excitación $\hat{e}[n] \approx 0$ para $n < N$ asumiendo, esto se podrá recuperar aproximadamente $\hat{h}[n]$ y $\hat{e}[n]$ de $\hat{x}[n]$.

El cepstrum $D[]$ de una señal digital $x[n]$ esta definido:

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln |X(e^{jw})| e^{jwn} \quad (3.47)$$

El análisis de \hat{x} nos permitirá conocer información del tracto vocal que se encuentra en la parte baja del cepstrum y la información del filtro (glotis) contenida en la parte

alta del cepstrum, luego se puede separar fácilmente $e[n]$ de $h[n]$ asumiendo el valor N antes mencionado y haciendo la operación inversa $D[]^{-1}$ del cepstrum.

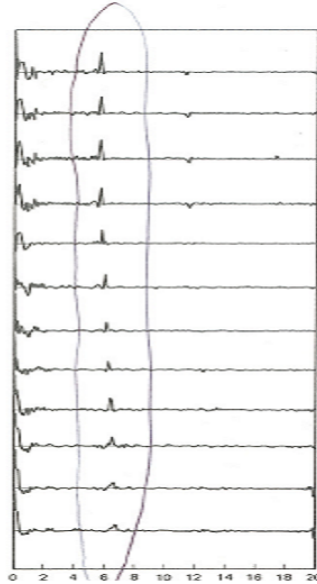


Figura 3.11 Coeficientes Cepstrales de la señal de habla, la parte baja corresponde al tracto vocal, la parte alta corresponde a la información provenientes de las cuerdas vocales. Fuente: Oppenheim

3.5.4 Extracción de características basadas en la Transformada de Fourier

Un método de extracción de características basado en un análisis de la señal mediante la Transformada de Fourier se hace aplicando la Transformada Corta de Fourier, usualmente utilizando una función del tipo Hamming para analizar segmentos de habla, luego se sacan las energías de diversos bandos de frecuencia o se hallan los bandos de frecuencia más significativos en la señal, estos valores constituirán las características de la señal de habla que serán los parámetros del reconocedor

3.5.5 Coeficientes Cepstrales en Frecuencia Mel

Un método más eficiente para sacar características y que es el más usado actualmente en reconocedores comerciales son los Coeficientes Cepstrales en Escala Mel, este método es un método robusto que hace uso de la Transformada de Fourier para obtener las frecuencias de la señal. El objetivo es desarrollar un conjunto de valores de características basados en criterios perceptuales, diversos experimentos muestran que la percepción de los tonos en los humanos no está dada una escala lineal, esto hace que se trate de aproximar el comportamiento del sistema auditivo humano.

Los Coeficientes Cepstrales en Frecuencia Mel (MFCC) son una representación definida como el cepstrum de una señal ventaneada en el tiempo que ha sido derivada de la aplicación de una Transformada Rápida de Fourier, pero en una escala en frecuencias no lineal, las cuales aproximan el comportamiento del sistema auditivo humano.

[Davis and Mermelstein, 1980] mostraron que los MFCC son beneficiosos para el Reconocimiento Automático del Habla.

Dada una Transformada Discreta de Fourier de una señal de entrada:

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi nk}{N}}, \quad k = 0, 1, \dots, N-1 \quad (3.48)$$

Se define un banco de filtros M , con $(m = 1, 2, \dots, M)$ donde el filtro m es un filtro triangular dado por:

$$H_m[k] = \begin{cases} 0 & \text{si } k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \leq k \leq f(m) \\ \frac{k-f(m)}{f(m+1)-f(m)} & f(m) \leq k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \quad (3.49)$$

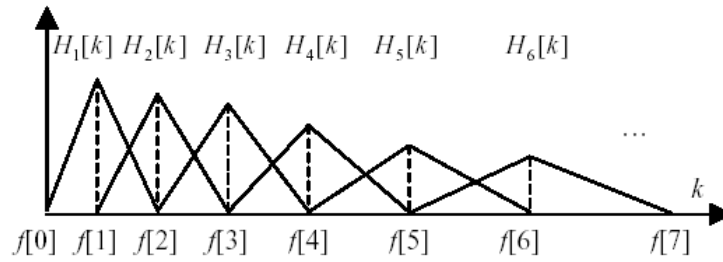


Figura 3.12 Filtros triangulares usados en el cálculo del Mel-Cepstrum.

Estos filtros calculan el promedio del espectro alrededor de cada frecuencia central.

Definimos f_l como la frecuencia más alta y f_h como la frecuencia más baja del banco de filtros en Hz, F_s es la frecuencia de Muestreo en Hz, M el número de filtros y N el tamaño de la Transformada Rápida de Fourier. Los puntos límite $f(m)$ son uniformemente espaciados en la escala Mel:

$$f(m) = \frac{N}{F_s} \beta^{-1} \left(\beta(f_1) + m \frac{\beta(f_h) - \beta(f_1)}{M+1} \right) \quad (3.50)$$

donde la escala Mel β esta dada por:

$$\beta(f) = 1125 \ln \left(1 + \frac{f}{700} \right) \quad (3.51)$$

y su inversa β^{-1} esta dada por:

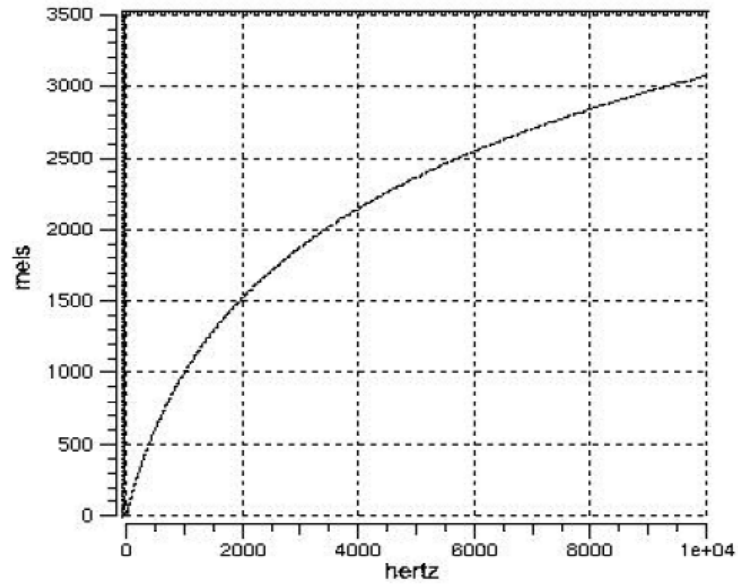


Figura 3.13 Escala perceptual Mel comparada con la escala de frecuencias.

$$\beta^{-1}[b] = 700(\exp(\frac{b}{1125}) - 1) \quad (3.52)$$

Entonces finalmente se computa el logaritmo de la energía de cada filtro:

$$S(m) = \ln(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)), \quad 0 < m < M \quad (3.53)$$

El Cepstrum en Frecuencia Mel es la Transformada Discreta del Coseno de las salidas de los M filtros:

$$c(m) = \sum_{n=0}^{M-1} S(n) \cos(\pi n(\frac{m - \frac{1}{2}}{M})) \quad (3.54)$$

donde M varía para diferentes implementaciones de 24 a 40, para el Reconocimiento Automático del Habla generalmente son usados los primeros 13 coeficientes.

Este algoritmo es ampliamente usado para obtener el vector de características en sistemas de Reconocimiento Automático del Habla.

Se puede también construir vectores de características basados en el modelo acústico de la producción del habla como es la técnica de Codificado de Predicción Lineal (LPC), también conocido como análisis autoregresivo y la técnica de Predicción Lineal Perceptual que hace uso del LPC y de los MFCC's.

3.6 Muestreo de la Señal de Habla

Para poder usar los métodos del procesamiento digital de la señal primeramente debemos digitalizar la señal de habla, que es una señal analógica, debiendo tomar periódicamente muestras de la señal analógica $x(t)$, a intervalos de T segundos para obtener la señal $x[t]$.

$$x[n] = x(nT) \quad (3.55)$$

donde T es el periodo de muestreo, luego $F_s = \frac{1}{T}$ es la frecuencia de muestreo, la cual para obtener buenos resultados debe estar por encima de los $8000Hz$, que es la frecuencia de las aplicaciones telefónicas.

Para poder capturar una frecuencia deseada, la frecuencia de muestreo F_s debe ser el doble de la frecuencia mas alta presente en la señal F , según el teorema de Nyquist.

$$F_s = 2 * F \quad (3.56)$$

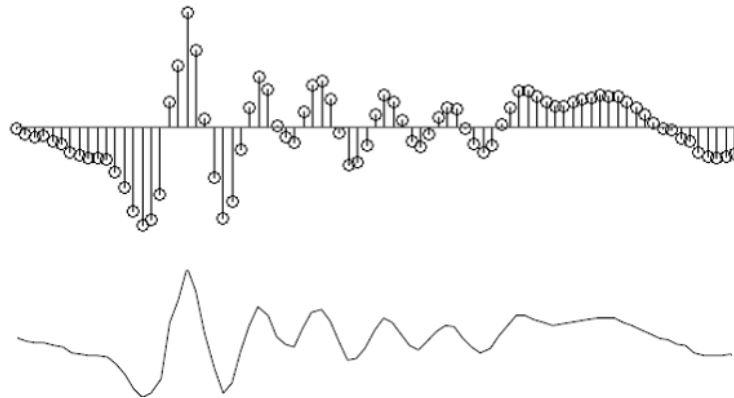


Figura 3.14 Muestreo de una señal de habla.

3.7 Codificación de la Señal de Habla

La señal de habla para que sea usada en la computadora necesita ser codificada digitalmente, el almacenamiento digital de una señal de habla puede resultar de alta calidad y de un tamaño pequeño comparado con su contraparte analógica, siendo posible almacenarlos en discos compactos, discos de video digital, MP3's, etc.

Las señales de audio de alta fidelidad son muestreadas a 44.1 KHz, el habla por teléfono es muestreada típicamente a 8 KHz, por eso las señales en teléfono están entre 300 y 3400 Hz, lo cual conlleva que el habla por teléfono tenga mala calidad. La reducción en la tasa de bits es uno de los propósitos principales del codificado del habla.

La complejidad computacional y requerimientos de memoria para el codificador de habla determinan el costo y poder del hardware en el que es implementado, en mu-

chos casos operaciones en tiempo real son requeridas al menos para el decodificador. Codificadores de habla pueden ser implementados en baratos procesadores digitales de señales que forman parte de muchos dispositivos, como máquinas contestadoras o DVD's donde el almacenamiento es más importante que el poder de cómputo; los procesadores digitales de señales también son usados en teléfonos celulares.

Podemos agrupar los codificadores en cuatro grupos:

- Codificadores Escalares de Forma de Onda.
- Codificadores Escalares en el Dominio de la Frecuencia.
- Codificadores Excitados de Predicción Lineal.
- Codificadores de Habla con Tasa Bajo-Bit.

3.7.1 Codificadores Escalares de Forma de Onda

Codifican cada muestra utilizando quantización escalar, estas técnicas intentan aproximar la forma de onda, mientras mayor sea el número de bits que se utilice para representar la señal, la señal codificada será máas parecida a la señal original; entre algunas técnicas de codificado de forma de onda tenemos: Linear Pulse Code Modulation (PCM), μ -law y A-law PCM, Adaptative PCM; Quantización Diferencial que puede ser: Differential Pulse Code Modulation (DPCM), Delta Modulation (DM), Adaptive Delta Modulation (ADM) llamada también Continuously Variable Slope Delta Modulation (CVSDM), Adaptive Differential PCM(ADPCM).

3.7.2 Codificación PCM

Es un Codificador Escalar de Forma de Onda, los convertidores analógicos-digitales hacen el proceso de muestreo y cuantización simultáneamente. Con B bits es posible representar 2^B niveles separados de cuantización, luego la salida del cuantificador $\hat{x}[n]$ está dada por:

$$\hat{x}[n] = Q\{x[n]\} \quad (3.57)$$

Muchos de los archivos de audio almacenados en las computadoras hacen uso de codificación PCM, tal es el caso de los archivos con formato Windows WAV, Apple AIF, Sun AU y SND entre otros formatos que usan codificación de 16-bits PCM; los discos compactos también usan codificación en 16-bits PCM.

Capítulo 4

Wavelets

Las wavelets han sido introducidos recientemente a principios de los años ochenta y han llegado a ser de gran interés en diversas disciplinas, pero sus raíces datan de mucho tiempo atrás.

4.1 Introducción

En la actualidad las wavelets han tomado una enorme popularidad. Sin embargo, sus raíces datan de 1873, cuando el trabajo de Karl Weierstrass describió una familia de funciones que son construidas por una superposición de copias escaladas de una función base dada. Las funciones que él definió son fractales, en el sentido de que son continuas en todos sus puntos y diferenciables en ninguno.

Otro trabajo importante fue el de Alfred Haar en 1909, cuando construyó el primer sistema ortonormal de funciones con soporte compacto, ahora llamada base de Haar. Esta base aún sirve como el fundamento para la moderna teoría del wavelet.

Otro avance significativo se dió en 1946, cuando Dennis Gabor describió una base no-ortogonal de lo que ahora se llaman wavelets con soporte no acotado, basado en funciones gaussianas trasladadas.

El término wavelet proviene del campo de la sismología, donde fue bautizado por Ricker en 1940 para describir el disturbio resultante de un impulso sísmico agudo o

una carga explosiva. En 1982, Morlet mostró como estos wavelets sísmicos podían ser modelados eficientemente con las funciones matemáticas que Gabor definió.

Posteriormente, Grossman y Morlet mostraron cómo señales arbitrarias pueden ser analizadas en términos de escalamientos y traslaciones de una función wavelet madre. Yves Meyer y Stephane Mallat ampliaron esta noción a una teoría llamada "análisis multiresolución". En 1989 Mallat mostró cómo esta teoría se puede utilizar en el procesamiento de imágenes y en el análisis de señales.

Una definición bastante sencilla de las wavelets es la que da Subhasis Saha: "Las wavelets son funciones definidas en intervalos finitos que tienen un valor promedio de cero". Las wavelets se definen por medio de una o varias funciones iniciales, llamadas wavelets madre, y un algoritmo para obtener el resto de las funciones que conformarán la base a partir de las funciones madre (algunos se refieren a este algoritmo como wavelet padre)

Estas bases de funciones han sido aplicadas a distintas áreas, como las ecuaciones diferenciales parciales, la física, el procesamiento de señales y gráficas computarizadas.

En el presente trabajo orientaremos las wavelets hacia el procesamiento de una señal de habla, que nos permitirán extraer información importante, que luego servirá para construir los patrones de características, los cuales nos permitirán posteriormente utilizarlos en un sistema de reconocimiento.

4.2 Transformada Wavelet

La Transformada Wavelet es una herramienta matemática que "corta" los datos, funciones o operadores en diferentes componentes de frecuencia [Daubechies, 1992] y estudia cada componente a una resolución ubicada a esa escala.

La justificación del uso de Wavelets se basa en el hecho de que la Transformada de Fourier no tiene una buena resolución Tiempo - Frecuencia, pues al analizar la señal lo trata como un todo y finalmente muestra las frecuencias presentes a lo largo de toda la señal, pero no indica el tiempo en que éstas ocurrieron, una alternativa es usar la Transformada Corta de Fourier, o llamada también la Transformada Ventaneada de Fourier, pero la desventaja de esta técnica es el escoger el tamaño de ventana que se sobrepondrá a la señal, si se coloca una ventana muy grande, se logra una buena resolución en frecuencias mas no en el tiempo, pues los picos y cambios instantáneos de frecuencia en la señal pasarían desapercibidos, en caso de notarlos, no se sabría el momento exacto donde ocurrieron; por el contrario si se escoge una ventana pequeña, se pierde resolución en la frecuencia, ganando resolución en el tiempo, acá las frecuencias muy bajas no se pondrían determinar con facilidad.

Una solución para este problema es la aplicación de las Wavelets que tiene buena localización Tiempo - Frecuencia de la señal y una justificación fisiológica en el sistema auditivo humano; en el oído interno, se encuentra el caracol llamado también cóclea, encargado de recibir la información proveniente del movimiento de las moléculas de

aire que hacen vibrar el tímpano, el que a su vez mueve unos huesecillos ubicados en el oído medio, transmitiendo de esta manera los sonidos percibidos hacia la cóclea, cuando los huesecillos estimulan la cóclea, se genera un movimiento en el líquido interior de ésta que a la vez excita los cilios que son como una especie de pilosidades, es en la cóclea donde se obtienen las frecuencias de la señal, captando las altas frecuencias, en la parte más externa y las bajas frecuencias en la parte más interna que contiene menos cilios; las Wavelets hacen algo parecido, al igual que la cóclea utiliza más cilios para analizar las altas frecuencias y menos cilios para analizar las bajas frecuencias, las Wavelets utilizan ventanas grandes para analizar la información de bajas escalas y ventanas pequeñas para analizar la información de altas escalas.

4.3 Transformada Wavelet Continua

Restringiendo a una dimensión y estableciendo los parámetros de dilatación y traslación a y b que varían continuamente sobre \Re con la restricción de $a \neq 0$, la transformada wavelet continua de una función f está dada por:

$$(T^{wav} f)(a, b) = |a|^{-\frac{1}{2}} \int \delta t f(t) \psi\left(\frac{t-b}{a}\right) \quad (4.1)$$

la familia de wavelets se puede construir dilatando y trasladando

$$\psi^{a,b}(x) = |a|^{-\frac{1}{2}} \psi\left(\frac{x-b}{a}\right) \quad (4.2)$$

entonces la transformada wavelet continua respecto a esta familia de wavelets es:

$$\begin{aligned}
(T^{wav} f)(a, b) &= \langle f, \psi^{a,b} \rangle \\
(T^{wav} f)(a, b) &= \int \delta x f(x) |a|^{-\frac{1}{2}} \psi\left(\frac{x-b}{a}\right)
\end{aligned} \tag{4.3}$$

la función f puede ser recuperada de su transformada wavelet como sigue:

$$f = C_{\psi}^{-1} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{\delta a \delta b}{a^2} \psi(T^{wav} f)(a, b) \psi^{a,b} \tag{4.4}$$

4.4 Comparación de la Transformada de Fourier con la Transformada Wavelet

A continuación se muestran las diferencias y similitudes del análisis Wavelet frente al análisis con Fourier.

$$T^{win}(w, t) = \int \delta s f(s) g(s-t) e^{-i\omega s} \tag{4.5}$$

Transformada Ventaneada de Fourier.

$$CWT(a, b) = \frac{1}{\sqrt{a}} \int x(t) \psi\left(\frac{t-b}{a}\right) \delta t \tag{4.6}$$

Transformada Wavelet. La transformada Wavelet provee una descripción similar Tiempo - Frecuencia. Una similitud entre la Transformada Wavelet y la Transformada Ventaneada de Fourier sería en que ambas toman el producto interno de la función f con una familia de funciones $g(s-t)e^{-i\omega s}$ y con $\psi\left(\frac{t-b}{a}\right)$, donde las funciones $\psi^{a,b}$ son llamadas Wavelets, siendo la función ψ llamada wavelet madre:

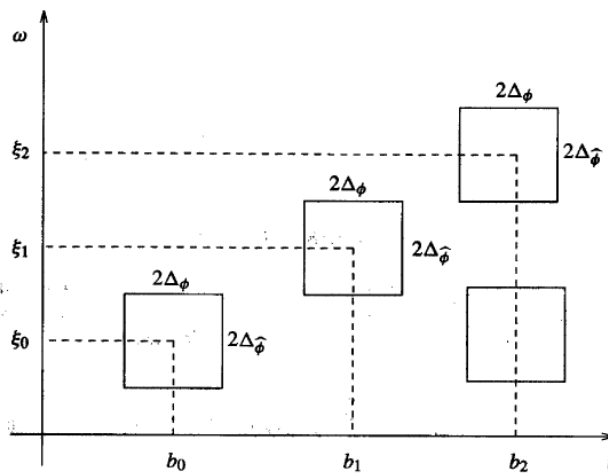


Figura 4.1 Ventana Tiempo-Frecuencia para la Transformada de Fourier. Fuente: [Mallat, 1989]

$$\psi^{a,b}(s) = |a|^{-\frac{1}{2}} \psi\left(\frac{s-b}{a}\right) \quad (4.7)$$

La diferencia entre la Transformada Wavelet y la Transformada Ventaneada de Fourier está dada en el hecho de la manera en como analizan las funciones [Daubechies, 1992], la función g analiza utilizando la misma forma para las frecuencias altas y las frecuencias bajas, la función ψ analiza las altas frecuencias con pequeñas formas y las bajas frecuencias con tamaño mucho mayores.

Algunas propiedades de las wavelets son:

condición de admisibilidad

$$\int \frac{|\psi(\omega)|^2}{|\omega|} < +\infty \quad (4.8)$$

esto se interpreta como que las wavelets se puede usar para reconstruir una señal sin

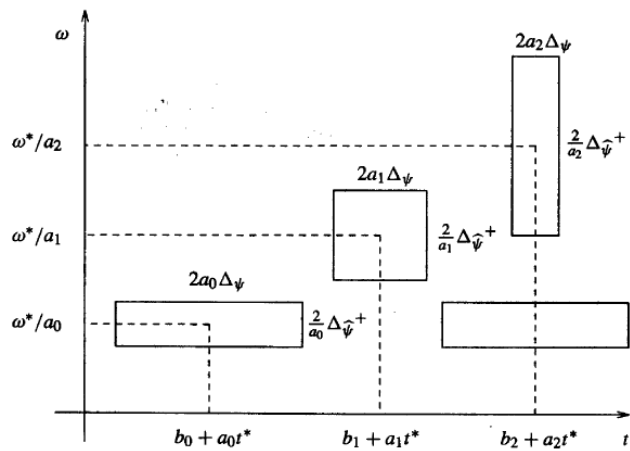


Figura 4.2 Ventana Tiempo-Frecuencia para la Transformada Wavelet. Fuente: [Mallat, 1989]

pérdida de información.

También implica lo siguiente:

$$|\psi(\omega)|^2|_{\omega=0} = 0 \quad (4.9)$$

Esto significa que las wavelets deben tener espectros parecidos a filtros pasa-banda, también significa que el valor promedio de las wavelets en el dominio del tiempo debe ser cero.

$$\int \psi(t) \delta t = 0 \quad (4.10)$$

esto quiere decir que ψ debe tener un comportamiento oscilatorio es decir debe ser una onda.

4.5 Transformada Wavelet Discreta

En este caso los parámetros de dilatación y traslación a, b , ambos toman solamente valores discretos, para a se escoge $a = a_0^m$ con $a_0 > 1$, diferentes valores de m corresponden a wavelets con diferentes anchos, entonces es de esperarse que la discretización del parámetro de traslación b dependa de m , en las frecuencias altas las wavelets son trasladados en pasos pequeños y en frecuencias bajas las wavelets son trasladados en pasos grandes, a fin de cubrir todo el rango del tiempo de la señal, si el ancho del wavelet es proporcional a a_0^m , entonces la discretización de $b = nb_0a_0^m$, donde b_0 es fijo y $n \in \mathbb{Z}$, luego las wavelets quedarán discretizados de la siguiente forma:

$$\begin{aligned}\psi^{m,n}(x) &= a_0^{-\frac{m}{2}} \psi(a_0^{-m}(x - nb_0a_0^m)) \\ \psi^{m,n}(x) &= a_0^{-\frac{m}{2}} \psi(a_0^{-m}x - nb_0)\end{aligned}\tag{4.11}$$

en particular si escogemos $a_0 = 2$ y $b_0 = 1$ entonces:

$$\psi^{m,n}(x) = 2^{-\frac{m}{2}} \psi(2^{-m}x - n)\tag{4.12}$$

Esta familia de funciones es llamada el set de expansión wavelet. La wavelet madre ψ , trae siempre asociada consigo una función escala ϕ . Con estas dos funciones podremos aproximar cualquier función o señal $f \in L^2$, mediante una de las funciones o mediante ambas, de la forma:

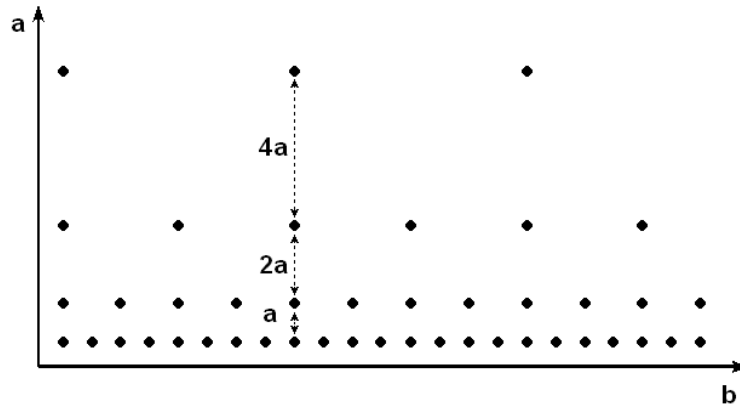


Figura 4.3 Localización de las wavelets discretos en el espacio tiempo-escala en una malla diádica. Fuente: [Mallat, 1989]

$$f(t) = \sum_m \sum_n c_{m,n} \phi(t) + \sum_m \sum_n d_{m,n}(t) \psi \quad (4.13)$$

La expansión wavelet entrega una localización tiempo-frecuencia instantánea de la señal, representación que puede explicarse como un pentagrama musical, donde la localización y forma de la figura musical nos dice cuando ocurre el tono y cual es su frecuencia. Esto quiere decir que la mayor parte de la energía de la señal es bien representada por unos pocos coeficientes. Un coeficiente de expansión wavelet representa un componente bien definido en un intervalo de tiempo.

Los sistemas wavelet satisfacen las condiciones de multi-resolución. Esto significa que si un conjunto de señales puede ser representado por una suma de $\phi(t - n)$ donde $n \in \mathbb{Z}$, un conjunto más amplio de señales (que incluye el conjunto original) puede ser representado por una suma $\phi(2t - n)$, $n \in \mathbb{Z}$.

Los coeficientes de más baja resolución pueden ser calculados a partir de los coeficientes de más alta resolución, mediante un algoritmo en forma de árbol, llamado banco de filtros. Esto permite un cálculo muy eficiente de los coeficientes de expansión.

El tamaño de los coeficientes de expansión wavelet disminuye rápidamente conforme aumenta m y n .

Las wavelets son ajustables y adaptables, debido a que existen muchas wavelets, estas pueden ser diseñados para adaptarse a una aplicación particular.

Las wavelets deben ser ortonormales esto es:

$$\int \psi_{j,k}(t)\psi_{m,n}^*(t)\delta t = \begin{cases} 1 & \text{si } j = m \text{ y } k = n \\ 0 & \text{caso contrario} \end{cases} \quad (4.14)$$

4.5.1 Función Escala

Sea $\phi \in L^2(\mathbb{R})$, una función escala, que trasladada y escalada genera una familia de funciones $\phi_{m,n} | m, n \in \mathbb{Z}$ definida como:

$$\phi_{m,n}(t) = 2^{-\frac{m}{2}} \phi(2^{-m}t - n) \quad (4.15)$$

$\forall m \in \mathbb{Z}$, la función escala define un espacio $V_m \subset L^2$ como:

$$V_m = \overline{\text{span}_{n \in \mathbb{Z}} \{ \phi_{m,n}(t) \}} \quad (4.16)$$

entonces la función $f(t)$ estará en V_m si puede escribirse como:

$$f(t) = \sum_{n \in Z} C_{m,n} \phi_{m,n}(t) \quad (4.17)$$

con

$$C_{m,n} = \langle f(t), \phi_{m,n}(t) \rangle \quad (4.18)$$

una propiedad importante de la función escala es:

$$f(t) \in V_m \iff f\left(\frac{t}{2}\right) \in V_{m+1} \quad (4.19)$$

Algunas características de la función escala son las siguientes:

- Los espacios V_m están anidados o sea $\forall j \in Z, V_m \subset V_{m-1}$
- La función ϕ tiene soporte compacto, es decir, existe un subconjunto del dominio de ϕ donde esta no es cero

4.5.2 Función Wavelet

Si se define

$$W = \overline{\text{span}_{n \in Z} \psi_{m,n}} \quad (4.20)$$

como el complemento ortogonal de V_m en V_{m-1} , esto significa que todos los miembros de V_m son ortogonales a todos los miembros de W_m . Entonces se requiere que:

$$\langle \phi_{m,n}, \psi_{m,n} \rangle = 0 \quad (4.21)$$

con

$$\psi_{m,n}(t) = 2^{\frac{-n}{2}} \psi(2^{-m}t - n) \quad (4.22)$$

y además

$$V_{m-1} = V_m \oplus W_m \quad (4.23)$$

donde cualquier función $f(t) \in W_m$ puede ser representado por:

$$f(t) = \sum_{k \in \mathbb{Z}} d_{m,n} \psi_{m,n}(t) \quad (4.24)$$

siendo la función wavelet básica:

$$\psi(t) = \psi_{0,0} \in W_0 \quad (4.25)$$

además

$$d_{m,n} = \langle f(t), \psi_{m,n}(t) \rangle \quad (4.26)$$

4.6 Análisis Multiresolución

El análisis Multiresolución forma el bloque más importante para la construcción de funciones escala y wavelets, el desarrollo de algoritmos es como el nombre sugiere un análisis multiresolución de una función vista a varios niveles de resolución, la idea fue desarrollada por Meyer [Meyer, 1986] y Mallat [Mallat, 1989]. Para aplicar el Análisis Multiresolución nosotros podemos dividir una función complicada en varias funciones simples y estudiarlas separadamente.

Adaptando la resolución de la señal, nos permite procesar solamente los detalles relevantes para cada tarea

La aproximación de una función en una resolución de 2^{-m} está definida como una proyección ortogonal en el espacio $V_m \in L^2(\mathfrak{R})$, el espacio V_m reagrupa todas las posibles aproximaciones a la resolución 2^{-m}

La proyección ortogonal de f es la función $f_m \in V_m$ que minimiza $\|f - f_m\|$

Definición 4.6.1 Una sucesión $\{V_m\}_{m \in \mathbb{Z}}$ de subespacios cerrados de $L^2\{\mathfrak{R}\}$ es un análisis multiresolución si las siguientes propiedades son satisfechas

$$\forall (m, n) \in \mathbb{Z}^2, \quad f(t) \in V_m \Leftrightarrow f(t - 2^m n) \in V_m \quad (4.27)$$

$$\forall m \in \mathbb{Z}, \quad V_{m+1} \subset V_m \quad (4.28)$$

$$\forall m \in \mathbb{Z}, \quad f(t) \in V_m \Leftrightarrow f\left(\frac{t}{2}\right) \in V_{m+1} \quad (4.29)$$

$$\lim_{m \rightarrow +\infty} V_m = \bigcap_{m=-\infty}^{+\infty} V_m = \{0\} \quad (4.30)$$

$$\lim_{m \rightarrow -\infty} V_m = \overline{\bigcup_{m=-\infty}^{+\infty} V_m} = L^2(\mathfrak{R}) \quad (4.31)$$

Existe ϕ tal que $\{\phi(t - n)\}_{n \in \mathbb{Z}}$ es una base de V_0

Como consecuencia de esta definición tenemos que un análisis multiresolución consiste en que los subespacios V_m satisfacen:

$$\dots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \dots \quad (4.32)$$

Si la resolución 2^{-m} tiende a 0, no tenemos ningún detalle de f .

$$\lim_{m \rightarrow +\infty} \|P_m f\| = 0 \quad (4.33)$$

, donde P_m es la proyección ortogonal en V_m

Por el contrario si la resolución 2^{-m} tiende a $-\infty$ significa que la aproximación de la señal tiende a ser la original.

$$\lim_{m \rightarrow -\infty} \|f - P_m f\| = 0 \quad (4.34)$$

Finalmente se requiere que exista $\phi \in V_0$ donde $\forall m, n \in \mathbb{Z}$, $\phi_{m,n}(x) = 2^{\frac{-m}{2}} \phi(2^{-m}x - n)$, esto implica que $\{\phi_{0,n}; n \in \mathbb{Z}\}$ es una base ortonormal para V_m $\forall m \in \mathbb{Z}$.

El principio básico del análisis multiresolución consiste en que si una colección de subespacios cerrados satisfacen las propiedades antes mencionadas, entonces existen bases ortonormales wavelets $\{\psi_{m,n}; m, n \in \mathbb{Z}\}$ de $L^2(\mathbb{R})$, $\psi_{m,n}(x) = 2^{\frac{-m}{2}} \psi(2^{-m}x - n)$, tal que para todo f en $L^2(\mathbb{R})$

$$P_{m-1}f = P_m f + \sum_{n \in \mathbb{Z}} \langle f, \psi_{m,n} \rangle \psi_{m,n} \quad (4.35)$$

Para cada $m \in \mathbb{Z}$ se define W_m como el complemento ortogonal de V_m en V_{m-1} entonces se tiene:

$$V_{m-1} = V_m \oplus W_m \quad (4.36)$$

esto implica

$$L^2(\mathfrak{R}) = \bigoplus_{m \in \mathbb{Z}} W_m \quad (4.37)$$

$$\forall m \in \mathbb{Z}, \quad f(t) \in W_m \Leftrightarrow f\left(\frac{t}{2}\right) \in W_{m+1} \quad (4.38)$$

4.7 Filtro Pasa Banda

Viendo las wavelets de otra manera es decir como un filtro pasa banda, entonces por análisis de Fourier se tiene que compresión en el tiempo es equivalente a estirar y trasladar el espectro hacia arriba (altas frecuencias).

$$H(f(at)) = \frac{1}{|a|} H\left(\frac{\omega}{a}\right) \quad (4.39)$$

Esto significa que una compresión en el tiempo del wavelet por un factor de 2 puede alargar el espectro por un factor de 2 y también trasladar los componentes de frecuencia por un factor de 2, usando esto podemos cubrir todo el espectro finito de nuestra señal con wavelets dilatados de la misma manera que podemos cubrir toda nuestra señal en el dominio del tiempo con wavelets trasladados.

Para tener un buen cubrimiento de la señal las wavelets trasladadas deberían tocarse uno con otro.

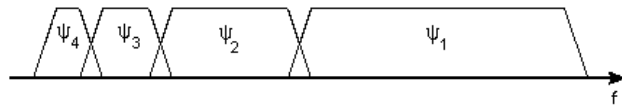


Figura 4.4 Localización de las wavelets discretas en el dominio de la frecuencia, como resultado de escalar las wavelets en el dominio del tiempo.

El radio de la frecuencia central del espectro de las wavelets y el ancho de su espectro, veremos que es el mismo para todos las wavelets, este radio es normalmente llamado factor de calidad Q y en el caso de las wavelets uno puede hablar de banco de filtros Q constante.

Para cubrir todo el espectro de la señal usaremos la función que trabajará como un espectro pasa baja, por la naturaleza de filtro pasa baja de la función escala también se le conoce como el filtro promedio.

Si vemos la función escala como justamente una señal, entonces podemos descomponerlo en componentes wavelet y expresarlo como sigue:

$$\phi(t) = \sum_{m,n} d_{m,n} \psi_{m,n} \quad (4.40)$$

Esto significa que si nosotros analizamos nuestra señal usando una combinación de funciones escala y wavelets, la función escala cubre el espectro de la señal que luego va a ser analizado con las wavelets a cierta escala m .

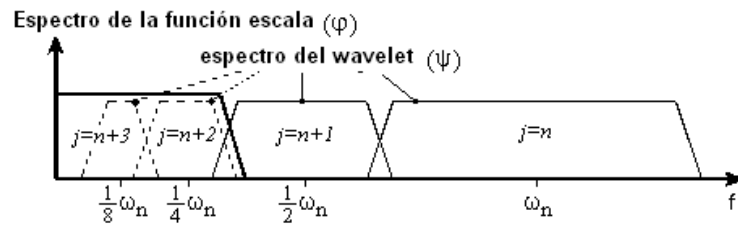


Figura 4.5 Función escala en el dominio de la frecuencia y como ésta es analizada por las wavelets.

4.8 Codificación de Subbanda

Viendo la Transformada Wavelet como un Banco de Filtros, entonces se considerará la Transformada Wavelet de una Señal como pasar una señal por esos bancos de filtros, la salida de diferentes estados de filtros son los coeficientes escala y coeficientes wavelet de la señal.

Al proceso de analizar la señal pasándola a través de un banco de filtros se conoce como codificación de subbanda, y tiene muchas aplicaciones como visión computacional, en nuestro caso el reconocimiento automático del habla.

El banco de filtros puede ser construido de muchas maneras, una manera es construir muchos filtros pasa banda y partir el espectro en muchos bandos de frecuencias, la ventaja de este método es que el ancho de cada bando de frecuencia puede ser escogido libremente, la desventaja es que se tiene que construir cada filtro separadamente y esto por ende mayor complejidad computacional.

Otra manera de construir el banco de filtros es partiendo el espectro de la señal

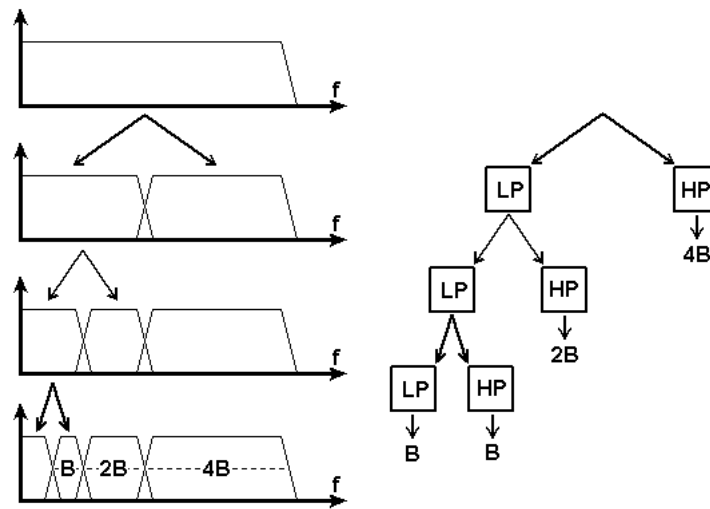


Figura 4.6 Banco de Filtros Iterativo.

en dos partes iguales: una parte pasa baja y una parte paso alto, la parte pasa alto contiene los pequeños detalles, la parte pasa baja contiene otros detalles que podemos analizar iterativamente partiendo el espectro de la parte pasa baja otra vez en dos partes iguales, y podemos seguir partiendo hasta un número suficientes de bandas que se necesite crear. De esta manera se ha construido un banco de filtros iterativo, la ventaja de este esquema es que solamente se han diseñado dos bancos de filtros, la desventaja es que el ancho del cubrimiento del espectro de la señal es fijo.

De este proceso se puede concluir que la Transformada Wavelet de una señal es lo mismo que la codificación subbanda de la señal, usando una constante Q banco de filtros, en general este tipo de análisis es el análisis multiresolución.

Desde que $\phi \in V_0 \subset V_{-1}$ y que $\phi_{-1,n}$ son bases ortonormales en V_{-1} , tenemos:

$$\phi(x) = \sqrt{2} \sum_n h_n \phi(2x - n) \quad (4.41)$$

con

$$h_n = \langle \phi, \phi_{-1,n} \rangle \quad (4.42)$$

y

$$\sum_{n \in \mathbb{Z}} |h_n|^2 = 1 \quad (4.43)$$

esto indica que la función escala en cierta nivel m puede ser expresada en términos de funciones escalas trasladadas en la siguiente escala mas pequeña.

Similarmente podemos expresar la función wavelet en cierto nivel en términos de funciones escaladas y trasladadas en la siguiente escala mas pequeña.

$$\psi(x) = \sqrt{2} \sum_n g_n \phi(2x - n) \quad (4.44)$$

con

$$g_n = \langle \psi, \phi_{-1,n} \rangle = (-1)^n h_{-n+1} \quad (4.45)$$

lo cual implica los dos importantes resultados:

Si una función f puede ser representada por funciones escala en el nivel m

$$f(t) = \sum_n C_n \phi_{-1,n} \quad (4.46)$$

con

$$C_n = \langle f, \phi_{-1,n} \rangle \quad (4.47)$$

y en términos de wavelets

$$f(t) = \sum_n d_n \psi_{-1,n} \quad (4.48)$$

con

$$d_n = \langle f, \psi_{-1,n} \rangle \quad (4.49)$$

finalmente se puede establecer lo siguiente:

$$C_m = \sum_l h(l - 2n)C_{m-1}(l) \quad (4.50)$$

$$d_m = \sum_l g(l - 2n)d_{m-1}(l) \quad (4.51)$$

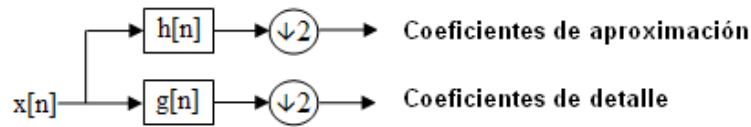


Figura 4.7 Esquema del banco de filtros.

Estas dos últimas ecuaciones nos dicen que los coeficientes wavelets y escala en cierto nivel m pueden ser encontrados de manera iterativa, por ejemplo empezando de $\langle f, \phi_{0,n} \rangle$ podremos calcular $\langle f, \psi_{1,n} \rangle$ y $\langle f, \phi_{1,n} \rangle$ y así sucesivamente.

Los valores para h y g actúan como filtros de paso baja y filtros de paso alto respectivamente y llamaremos a h el filtro escala y a g el filtro wavelet.

4.9 Complejidad Computacional de la Transformada Wavelet

La complejidad computacional de este esquema de filtros es la siguiente:

$$T[n] = \begin{cases} T(\frac{n}{2}) + Cn & \text{si } 2^n \geq 2 \\ 0 & n = 1 \end{cases} \quad (4.52)$$

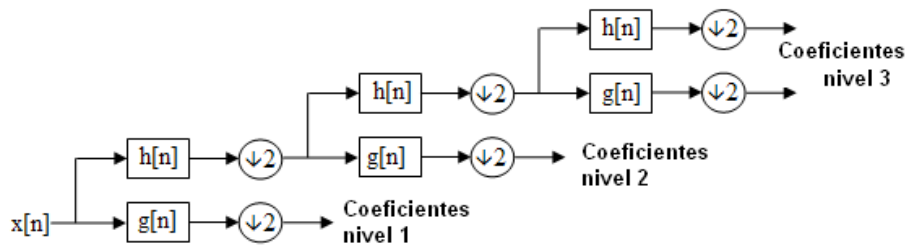


Figura 4.8 Implementación del banco de filtros iterativo.

resolviendo la ecuación de recurrencia

$$T(n) = T\left(\frac{n}{2}\right) + Cn$$

$$T\left(\frac{n}{2}\right) = T\left(\frac{n}{2^2}\right) + C\frac{n}{2} + Cn$$

$$\dots = \dots$$

$$2^a = n$$

$$a = \log_2 n$$

$$\dots = \dots$$

$$T\left(\frac{n}{2^{a-1}}\right) = T\left(\frac{n}{2^a}\right) + \underbrace{C\frac{n}{2^{a-1}} + C\frac{n}{2^{a-2}} + \dots + \dots + Cn}_{\dots}$$

$$T\left(\frac{n}{2^{a-1}}\right) = + (cn) \sum_{i=0}^{a-1} \frac{1}{2^i}$$

$$T(n) = cn \left(\frac{\frac{1}{2^a} - 1}{\frac{1}{2} - 1} \right)$$

$$T(n) = cn \left(\frac{\frac{1}{2} - 1}{-\frac{1}{2}} \right)$$

$$T(n) = cn \left(\frac{2n - 2}{n} \right)$$

$$T(n) = 2cn - 2c$$

La Transformada wavelet con banco de filtros tiene una complejidad de $O(n)$.

4.10 Wavelet Packets

Las Wavelets Packets fueron introducidos por Coifman, Meyer y Wickerhauser [Coifman Meyer and Wickerhauser, 1992], en lugar de hacer el procedimiento solo para los espacios de aproximación V_m para construir los espacios detalle W_m y las bases Wavelet, podemos también dividir los espacios detalle W_m y derivar nuevas bases, esto se puede representar como un árbol binario, donde el nodo raíz está asociado con el espacio de aproximación V ; todo nodo en el árbol, donde m es la profundidad del árbol y p es el número de hojas del árbol, cada nodo (m, p) tiene asociado un espacio W_m^p , cada cual admite bases ortonormales $\psi_m^p(t - 2^m n)_{n \in \mathbb{Z}}$

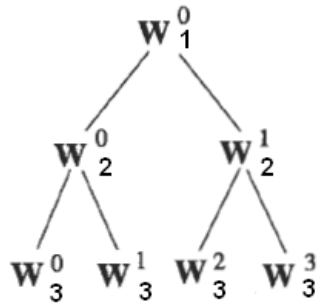


Figura 4.9 Árbol de descomposición para el Wavelet Packet, cada nodo forma un espacio W .

Haciendo el nodo raíz $W_m^0 = V$ y además $\psi_m^0 = \phi_0$, los dos nuevos wavelets packets

de los nodos hijos serán:

$$\psi_{m+1}^{2p}(t) = \sum_{n=-\infty}^{+\infty} h[n]\psi_m^p(t - 2^m n) \quad (4.53)$$

y

$$\psi_{m+1}^{2p+1}(t) = \sum_{n=-\infty}^{+\infty} g[n]\psi_m^p(t - 2^m n) \quad (4.54)$$

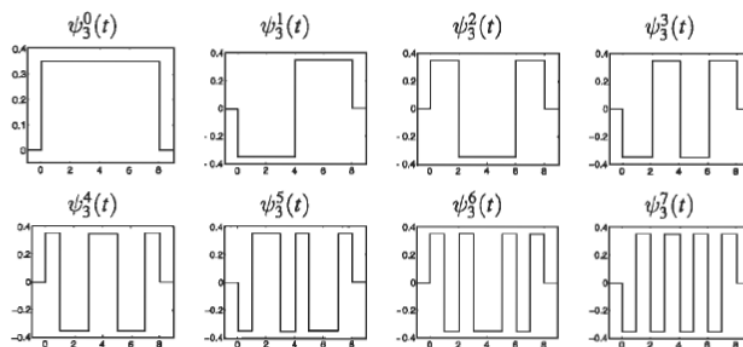


Figura 4.10 Wavelet Packet en profundidad 3, mejor cubrimiento de rango de frecuencias calculado con el Wavelet de Haar.

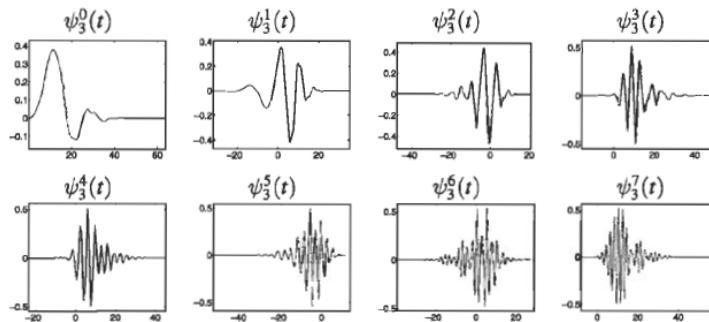


Figura 4.11 Wavelet Packet en profundidad 3, mejor cubrimiento de rango de frecuencias calculado con el Wavelet de Daubechies.

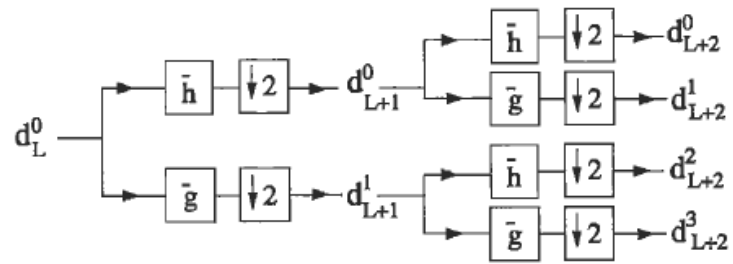


Figura 4.12 Cálculo de las Wavelets Packets mediante banco de filtros iterativo.

Capítulo 5

Técnicas para el Reconocimiento Automático del Habla

El Reconocimiento Automático del Habla puede ser visto como un problema de Reconocimiento de Patrones, existen técnicas útiles como son las Redes Neuronales Artificiales, Modelos Ocultos de Markov, etc.

Para la presente Tesis, se optó por hacer uso de la técnica llamada Dynamic Time Warping (DTW), pues es muy útil para construir reconocedores de palabras aisladas, que es suficiente para hacer las pruebas de los algoritmos de procesamiento de la señal basados en las wavelets frente a los basados en la Transformada de Fourier.

A continuación se muestra cada técnica de manera general.

5.1 Redes Neuronales Artificiales

Son modelos computacionales que tratan de imitar el comportamiento de las neuronas humanas, son buenos clasificadores y existen muchos modelos los cuales son aplicables para ciertos problemas en particular.

En el Reconocimiento Automático del habla han sido usadas redes como: la red neuronal de Kohonen, redes neuronales de clasificación espacio temporal de tramas, entre otras.

5.2 Modelos Ocultos de Markov

Son modelos estadísticos de caracterización de muestras de datos, los Modelos Ocultos de Markov proveen un eficiente camino para construir buenos modelos, también incorporan el principio de programación dinámica como su núcleo para hacer segmentación y clasificación de patrones de secuencias de datos.

Esta técnica fué publicada por: [Baun and Eagon, 1967] y ha llegado a ser desde entonces el método estadístico mas poderoso para modelar señales de habla.

5.3 Dynamic Time Warping

Consideremos el siguiente escenario: para cada palabra w obtenemos muestras de audio A_w , luego para reconocer una señal de entrada A , tenemos que buscar la palabra w que minimize alguna distancia $D(A, A_w)$, teniendo en cuenta de que para palabras iguales tendremos vectores de características similares.

Analizando tendremos algunos casos, como que la señal de audio de entrada A tenga la misma longitud que la señal de audio almacenada A_w , si es de esta manera, cada señal de audio nos proporcionar'a el mismo número de elementos en el vector de características resultante después de hacer el procesamiento digital de la señal, lo que se reduce simplemente a una medida de distancia:

$$D(A, A_w) = \sum_{t=1}^T DF(A(t), A_w(t)) \quad (5.1)$$

donde DF es la distancia entre frames, y T es el número de Frames que tiene una

señal de entrada, luego la evaluación de la distancia se podría hacer de varias formas:

La distancia euclidiana:

$$\sqrt{\sum_i (A_i - A'_i)^2} \quad (5.2)$$

Norma L^P :

$$\sqrt[p]{\sum_i |A_i - A'_i|^P} \quad (5.3)$$

entre otras distancias.

Pero si se tiene el caso de que la señal de audio de entrada A , y la señal de audio almacenada A_w no tienen el mismo tamaño, la manera mas fácil de solucionar este problema sería hacer una normalización lineal en el tiempo es decir:

$$D(A, A_w) = \sum_{t=1}^T DF(A(t), A_w(t)) \quad (5.4)$$

donde:

$$t' = t \frac{\text{longitud}(A_w)}{\text{longitud}(A)} \quad (5.5)$$

donde se hace una normalización lineal en el tiempo. La normalización lineal en el tiempo trabaja mal, pues en realidad solo hace coincidir dos palabras para que tengan la misma dimensión más no tiene en cuenta similitudes en los vectores de características, como por ejemplo que en una palabra se pronuncie una vocal en un tiempo prolongado y se haga una normalización en el tiempo con otra que no se pronunció de la misma manera, así tenemos la palabra "casa" dicha de la

siguiente manera c-a-a-s-s-a y la misma palabra dicha de diferente manera c-c-a-s-a-a-a, como se ve ambas tienen el mismo tamaño, que puede ser el resultado de hacer una normalización lineal en el tiempo, cuando se haga un cálculo de distancias de los vectores de características esto llevará a resultados erróneos.

5.3.1 Distancia General Normalizada en el Tiempo

Si definimos que una señal de audio, que en nuestro caso sería el habla, puede ser representada apropiadamente por su vector de características resultante del procesamiento de la señal tenemos:

$$A = a_1, a_2, \dots, a_i, \dots, a_I \quad (5.6)$$

$$B = b_1, b_2, \dots, b_j, \dots, b_J \quad (5.7)$$

Si eliminamos las diferencias en el tiempo de aquellos dos vectores de características de señales de habla, se puede considerar un plano bidimensional i, j , donde los patrones A y B se encuentran en los ejes i y j respectivamente, entonces si tenemos patrones de la misma categoría la alineación en el tiempo puede hacerse de la siguiente manera:

$$F = c(1), c(2), \dots, c(k), \dots, c(K) \quad (5.8)$$

donde:

$$c(k) = (i(k), j(k)) \quad (5.9)$$

es decir se realiza un mapeo del patrón A en el patrón B , a esto se llamará función warping. Si no hay diferencias en el tiempo entre los dos patrones, la función warping coincide con la línea diagonal $i = j$, si existieran diferencias en el tiempo, la función warping tendrá algún desvío de la diagonal.

Como medida de diferencia entre dos vectores de características a_i y b_j , una distancia:

$$d(c) = d(i, j) = \|a_i - b_j\| \quad (5.10)$$

entonces una suma con pesos de la función warping sería:

$$E(F) = \sum_{k=1}^K d(c(k)) \cdot w(k) \quad (5.11)$$

donde los $w(k)$, son coeficientes no negativos, los cuales son introducidos para permitir que $E(F)$ sea una medida de características flexibles, y esta medida indica una buena función warping; tiene su mínimo valor cuando la función warping está perfectamente ajustada.

La distancia normalizada entre dos patrones A y B está definida como:

$$D(A, B) = \min \left[\frac{\sum_{k=1}^K d(c(k)) \cdot w(k)}{\sum_{k=1}^K w(k)} \right] \quad (5.12)$$

donde el denominador $\sum_{k=1}^K w(k)$ es empleado para compensar el efecto de K , siendo K el número de puntos de la función warping F . Una condición para la aplicación de esta técnica es que ambos patrones a compararse estén muestreados

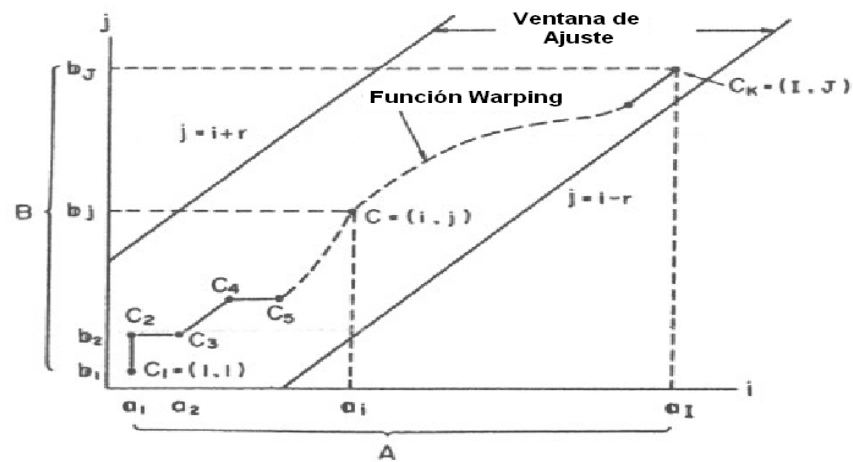


Figura 5.1 Función Warping y Ventana de Ajuste. Fuente: [Sakoe and Chiba, 1978]

con el mismo periodo de muestreo, y que no se tenga conocimiento a priori donde se encuentra la información importante en los patrones de habla.

5.3.2 Restricciones de la Función Warping

La función warping, es una medida de la fluctuación en el eje del tiempo de los patrones del habla, la función F puede ser vista como una función de mapeo del patrón A en el patrón B , quien debería conservar las estructuras lingüísticas esenciales en el patrón A y viceversa.

Las restricciones de la Función warping F , son las siguientes:

Condición de Monotonicidad:

$$i(k-1) \leq i(k) \quad (5.13)$$

y

$$j(k-1) \leq j(k) \quad (5.14)$$

Condición de Continuidad:

$$i(k) - i(k - 1) \leq 1 \quad (5.15)$$

y

$$j(k) - j(k - 1) \leq 1 \quad (5.16)$$

como resultado de estas restricciones, la relación entre dos puntos consecutivos está dada:

$$c(k - 1) = \begin{cases} (i(k), j(k - 1)) \\ (i(k - 1), j(k - 1)) \\ (i(k - 1), j(k)) \end{cases} \quad (5.17)$$

Condición de frontera:

$$i(1) = 1, \quad j(1) = 1 \quad (5.18)$$

y

$$i(K) = I, \quad j(K) = J \quad (5.19)$$

Condición de ventana de ajuste:

$$|i(k) - j(k)| \leq r \quad (5.20)$$

donde r es un apropiado valor entero no negativo que indica el tamaño de la ventana de ajuste, esto se debe al hecho en que las fluctuaciones en el eje del tiempo, en algunos casos nunca causan excesiva diferencia.

Condición de Slope Constraint: No se deben permitir gradientes ni muy pronunciadas, ni muy suaves para la función F , pues puede causar desviaciones no deseables en el eje del tiempo; para gradientes pronunciadas puede causar una correspondencia no real entre un patrón muy corto y otro muy largo.

La condición de slope constraint es establecida como la primera derivada de la función warping en su forma discreta, así se obliga que los valores $c(k)$ se muevan en dirección hacia adelante en el eje i o en el eje j , consecutivamente m veces, entonces no se debe permitir que $c(k)$ vaya en la misma dirección a menos que haya pasado n veces por la diagonal, la intensidad efectiva del slope constraint puede ser evaluada por la siguiente medida:

$$p = \frac{n}{m} \quad (5.21)$$

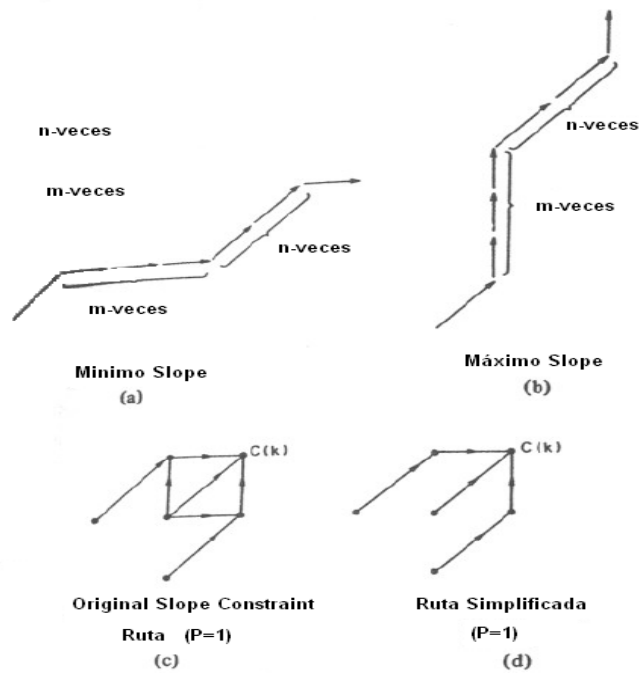


Figura 5.2 Slope Constraint en Función Warping. Fuente: [Sakoe and Chiba, 1978]

Cuando $P = 0$, no existen restricciones en la función warping, cuando $P = \infty$

que significa $m = 0$ la función warping esta restringida a la diagonal $i = j$, si el slope constraint es muy severo, entonces la normalización en el tiempo no podrá trabajar adecuadamente y si el slope constraint es muy relajado, entonces la discriminación entre patrones de diversas categorías es degradada, así es deseable un valor ni muy pequeño, ni muy grande de p .

5.3.3 Coeficientes de Pesos

La expresión

$$N = \sum_{k=1}^K w(k) \quad (5.22)$$

que actúa como denominador de el cálculo de la distancia normalizada:

$$D(A, B) = \min \left[\frac{\sum_{k=1}^K d(c(k)) \cdot w(k)}{\sum_{k=1}^K w(k)} \right] \quad (5.23)$$

es independiente de la función warping F , luego la distancia normalizada se puede escribir como:

$$D(A, B) = \frac{1}{N} \min \left[\sum_{k=1}^K d(c(k)) \cdot w(k) \right] \quad (5.24)$$

y puede ser efectivamente resuelto por la técnica de programación dinámica. Existen dos definiciones de coeficientes de pesos:

Forma simétrica:

$$w(k) = (i(k) - i(k - 1)) + (j(k) - j(k - 1)) \quad (5.25)$$

entonces $N = I + J$, donde I y J son las longitudes de los parámetros de habla A y B respectivamente.

Forma asimétrica:

$$w(k) = (i(k) - i(k - 1)) \quad (5.26)$$

entonces $N = I$, ó equivalentemente:

$$w(k) = j(k) - j(k - 1) \quad (5.27)$$

entonces $N = J$.

La forma simétrica significa que si los ejes del tiempo i y j son continuos, entonces existirá un eje temporal: $l = i + j$; en la forma asimétrica se refiere a la integración a través del eje i o j según sea el caso. Según [Sakoe and Chiba, 1978], se espera que trabaje mejor la forma simétrica que la asimétrica, pues trata las partes del vector de características de igual manera, y la forma asimétrica hace alguna exclusion cuando la función warping está en dirección del eje i ó j según sea el caso, pues $w(k)$ se reduce a cero: $c(k) = c(k - 1) + (0, 1)$, esto significa que algunos vectores pueden ser excluidos del análisis, pero por suerte la condición de slope constraint reduce esta situación, la diferencia de desempeño entre la forma simétrica y asimétrica puede ser gradualmente atenuada, del mismo modo que la condición de slope constraint es intensificada.

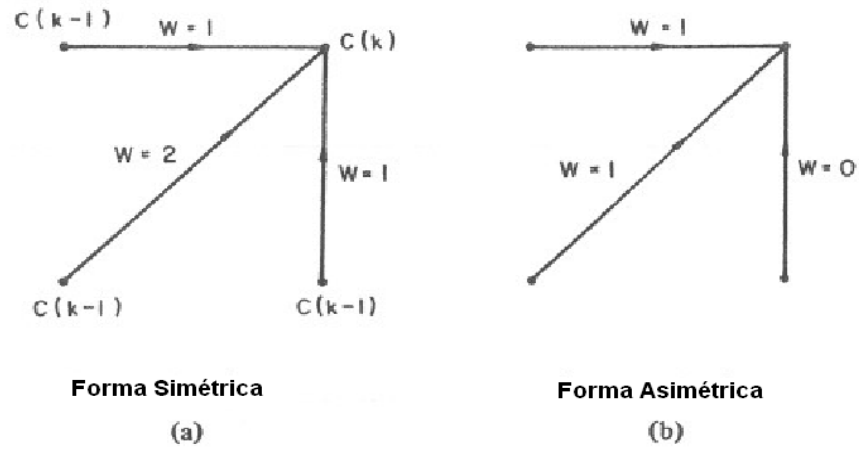


Figura 5.3 Coeficientes de pesos para la forma simétrica y la forma asimétrica. Fuente: [Sakoe and Chiba, 1978]

5.3.4 Algoritmo PD-Matching

El principio de Programación dinámica (PD) puede ser aplicado a la definición de distancia normalizada en el tiempo, el algoritmo básico es descrito como sigue:

Condición Inicial:

$$g_1(c(1)) = d(c(1))w(1) \quad (5.28)$$

Ecuaciones PD:

$$g_k(c(k)) = \min_{c(k-1)} [g_{k-1}(c(k-1)) + d(c(k))w(k)] \quad (5.29)$$

Distancia Normalizada en el Tiempo:

$$D(A, B) = \frac{1}{N} g_k(c(k)) \quad (5.30)$$

Se asume que $c(0) = (0, 0)$ y $w(1) = 2$ para la forma simétrica y $w(1) = 1$ para la forma asimétrica.

Pueden derivarse varios algoritmos prácticos al aplicar la forma simétrica y la forma asimétrica según la condición de Slope Constraint que utilicen, alguno de ellos se detallan a continuación:

Algoritmo de la forma Simétrica

Condición Inicial:

$$g(1, 1) = 2d(1, 1) \quad (5.31)$$

Ecuación PD: para $p = 0$ en su forma simétrica

$$g(i, j) = \min \left[\begin{array}{l} g(i, j - 1) + d(i, j) \\ g(i - 1, j - 1) + 2d(i, j) \\ g(i - 1, j) + d(i, j) \end{array} \right]. \quad (5.32)$$

para $p = 1$ en su forma simétrica será

$$g(i, j) = \min \left[\begin{array}{l} g(i - 1, j - 2) + 2d(i, j - 1) + d(i, j) \\ g(i - 1, j - 1) + 2d(i, j) \\ g(i - 2, j - 1) + 2d(i - 1, j) + d(i, j) \end{array} \right]. \quad (5.33)$$

Condición de restricción (ventana de ajuste)

$$j - r \leq i \leq j + r \quad (5.34)$$

Distancia Normalizada en el Tiempo:

$$D(A, B) = \frac{1}{N}g(I, J) \quad (5.35)$$

Las ecuaciones de PD ($g(i, j)$) se calculan en orden ascendente con respecto a las coordenadas i y j , es decir empiezan de (i, j) hasta (I, J)

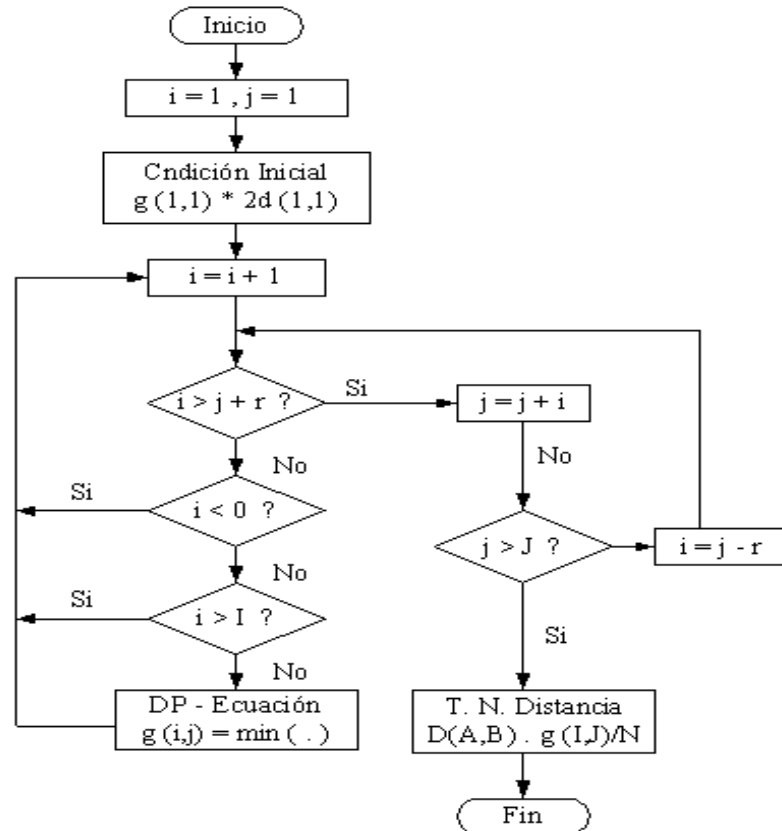


Figura 5.4 Diagrama de flujo del algoritmo PD-Matching.

En experimentos realizados por [Sakoe and Chiba, 1978] muestran que el desempeño de la forma simétrica es superior al de la forma asimétrica, pero esta diferencia disminuye a medida que la condición de slope constraint es intensificada, se puede notar también que la desempeño de la forma simétrica no es afectada por un slope con-

straint superior a $P = 1$, por otro lado la forma asimétrica es visiblemente mejorada por la condición de slope constraint. Otro experimento hecho también por [Sakoe and Chiba, 1978] evalúa el efecto de la condición de slope constraint en la forma simétrica del algoritmo PD-matching, y llega al resultado de que cuando $P = 1$ se obtiene la mejor desempeño, y en otro experimento evalúa este algoritmo con otros varios algoritmos PD propuestos por otros autores, y muestra la superioridad del algoritmo PD-matching con slope constraint $P = 1$.

En conclusión se tiene que los mejores resultados se obtienen con la forma simétrica del algoritmo pd-matching y con una condición de slope constraint $P = 1$.




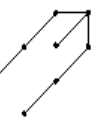
P	Esquema	Ecuación DP $g(i, j) =$	
		Simétrico Asimétrico	
0			$\min \begin{pmatrix} g(i, j-1) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i-1, j) + d(i, j) \end{pmatrix}$
			$\min \begin{pmatrix} g(i, j-1) \\ g(i-1, j-1) + d(i, j) \\ g(i-1, j) + d(i, j) \end{pmatrix}$
$\frac{1}{2}$		Simétrico	$\min \begin{pmatrix} g(i-1, j-3) + 2d(i, j-2) + d(i, j-1) + d(i, j) \\ g(i-1, j-2) + 2d(i, j-1) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i-2, j-1) + 2d(i-1, j) + d(i, j) \\ g(i-3, j-1) + 2d(i-2, j) + d(i-1, j) + d(i, j) \end{pmatrix}$
		Asimétrico	$\min \begin{pmatrix} g(i-1, j-3) + (d(i, j-2) + d(i, j-1) + d(i, j))/3 \\ g(i-1, j-2) + (d(i, j-1) + d(i, j))/2 \\ g(i-1, j-1) + d(i, j) \\ g(i-2, j-1) + d(i-1, j) + d(i, j) \\ g(i-3, j-1) + d(i-2, j) + d(i-1, j) + d(i, j) \end{pmatrix}$
1		Simétrico	$\min \begin{pmatrix} g(i-1, j-2) + 2d(i, j-1) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i-2, j-1) + 2d(i-1, j) + d(i, j) \end{pmatrix}$
		Asimétrico	$\min \begin{pmatrix} g(i-1, j-2) + (d(i, j-1) + d(i, j))/2 \\ g(i-1, j-1) + d(i, j) \\ g(i-2, j-1) + d(i-1, j) + d(i, j) \end{pmatrix}$
2		Simétrico	$\min \begin{pmatrix} g(i-2, j-3) + 2d(i-1, j-2) + 2d(i, j-1) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i-3, j-2) + 2d(i-2, j-1) + 2d(i-1, j) + d(i, j) \end{pmatrix}$
		Asimétrico	$\min \begin{pmatrix} g(i-2, j-3) + 2(d(i-1, j-2) + d(i, j-1) + d(i, j))/3 \\ g(i-1, j-1) + d(i, j) \\ g(i-3, j-2) + d(i-2, j-1) + d(i-1, j) + d(i, j) \end{pmatrix}$

Figura 5.5 Algoritmos simétricos y asimétricos con condición de Slope Constraint $P = 0, \frac{1}{2}, 1, 2$. Fuente: [Sakoe and Chiba, 1978]

Capítulo 6

Reconocimiento Automático del Habla utilizando Wavelets

Para el desarrollo de la presente tesis, se implementaron algoritmos basados en las wavelets y wavelets packets, para la construcción de estos últimos se tuvo en cuenta el rango de frecuencias perceptuales en la escala Mel, es por eso que se decidió llamarlo wavelet packet perceptual para la extracción de características,

El esquema a seguir para ambos modelos es el siguiente:

Primero se muestreó la señal analógica con una frecuencia de muestreo igual a $F_s = 16000Hz$, lo que nos permitió analizarla hasta un rango de frecuencias de $8000Hz$, la salida de este procedimiento es la señal digital.

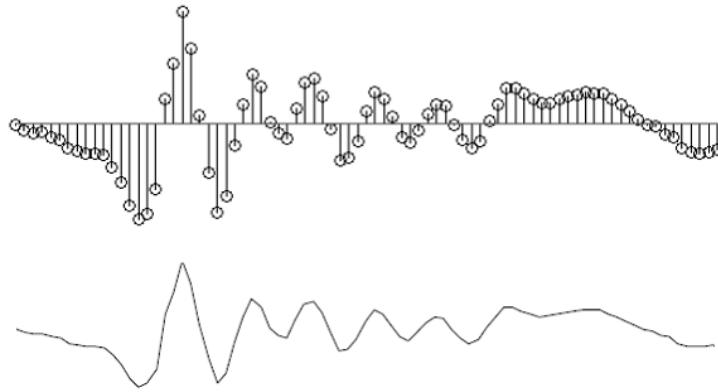


Figura 6.1 Muestreo de la señal analógica para construir la señal digital.

Luego se aplicó una cuantificación a la señal del tipo PCM con signo de 16 bits, esto

nos permitió tener un amplio rango de valores de amplitud ($2^{16} = 65536$ posibles valores), como la cuantificación fué del tipo PCM con signo, se obtuvieron valores entre -32768 y $+32767$ que representaban la amplitud de la señal.

Posteriormente se procedió a implementar un algoritmo de eliminación de segmentos inútiles, es decir eliminar las secciones de las señales que no corresponden a habla, tales como los valores de inicio y de final captados ya sea por la demora en pronunciar una palabra o en detener la grabación. Para ello se obtuvo un umbral, con el cual vamos a comparar los valores de las muestras, si existen valores que estén por debajo de éste, los valores se irán eliminando. El umbral es simplemente un valor promedio de los valores del principio y del final de la señal, éste umbral nos permite elegir el punto de corte de la señal.

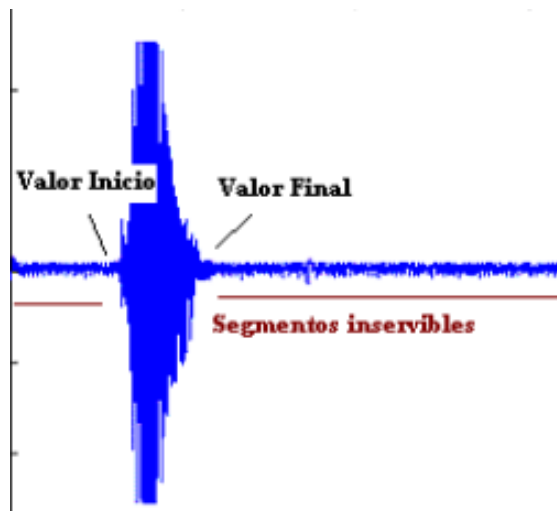


Figura 6.2 Obtención de los segmentos con información de la señal de habla.

Luego de recortar la señal, las señales resultantes tienen menos información no válida que las iniciales. Posteriormente se procedió a hacer un ventaneamiento de la señal con una ventana hamming, descrita en la sección 3.4, para la ventana hamming se estableció un tamaño de $16mS$ y de $32mS$ en otros casos, el tamaño de paso que se utilizó fue de $10mS$.

Esto conforme a lo siguiente:

$$x^m[n] = x[n - mF]w[n] \quad (6.1)$$

El tamaño de paso, estuvo aproximadamente entre $[10mS - 20mS]$ para poder obtener información consistente.

Una vez llegado a este punto se procedió a aplicar la Transformada Wavelet utilizando el algoritmo de banco de filtros para cada frame y también la transformada wavelet packet basada en las frecuencias perceptuales, los procedimientos se detallan a continuación.

6.1 Modelo propuesto para la Extracción de Características basadas en las wavelets

De la misma manera que los Coeficientes Cepstrales en Escala Mel descritos en la sección 3.5.5, hacen un filtrado de los componentes de frecuencia, luego una decorrelación del cepstrum mediante una transformada del coseno, la extracción de características en el procesamiento digital de la señal mediante wavelets hacen algo similar.

Primero se hizo una descomposición wavelet de cada frame hasta m niveles de descomposición que corresponderán a un análisis multiresolución, donde la señal es proyectada a cada nivel de resolución obteniendo al final m espacios W correspondientes a diversos rangos de frecuencia y un espacio V correspondiente al nivel mas bajo de frecuencia de la señal.

En el espacio V y en cada espacio W , tendremos coeficientes escala ϕ y coeficientes wavelet ψ correspondientes y por ende podremos reconstruir la señal a cada nivel de resolución, en este caso solo se utilizarán los coeficientes $C_{m,n}$ y $d_{m,n}$ para determinar información importante en determinado espacio de tiempo-frecuencia en la señal de habla, coeficientes con altos valores nos indicarán la presencia de información importante, esto dependerá del tipo de función wavelet utilizado, para la presente tesis se utilizaron las siguientes funciones wavelet: Haar, Daubechies 4, Daubechies 6, Coiflets 6, estas funciones wavelet escaladas y trasladadas por toda la señal de habla tratarán de encontrar partes en el tiempo donde son más parecidas a la señal original produciéndose valores altos para los coeficientes en determinado nivel y zonas donde la señal no tiene cambios aparentes produciéndose coeficientes con valores muy bajos.

A continuación se detallan los componentes de frecuencia aproximados por nivel de resolución

Para proceder a obtener la extracción de características mediante wavelets, se hizo un ventaneamiento de la señal con el procedimiento detallado en la sección 3.4.

nivel (m)	frecuencia (Hz) espacio V	frecuencia (Hz) espacio W
0	0hz-8000hz	0hz-8000hz
1	0hz-4000hz	4000hz-8000hz
2	0hz-2000hz	2000hz-4000hz
3	0hz-1000hz	1000hz-2000hz
4	0hz-500hz	500hz-1000hz
5	0hz-250hz	250hz-500hz
6	0hz-125hz	125hz-250hz

Tabla 6.1 Rango aproximado de frecuencias en los espacios V y W para una frecuencia de muestreo igual a 16000Hz y un nivel de descomposición $m = 7$.

Una vez obtenidos los valores del ventaneamiento se procedió a hacer una descomposición wavelet de cada segmento obtenido del ventaneamiento, hasta un nivel $m=7$ para las wavelets de Haar, $m=6$ para las wavelets de Daubechies 4, $m=5$ para las wavelets de Daubechies 6 y Coiflets 6.

El proceso de descomposición utilizó el algoritmo de banco de filtros que a continuación se detalla para cada tipo de wavelet de los cuales daremos los filtros correspondientes.

6.1.1 Wavelet de Haar

Este wavelet también conocido como wavelet de Daubechies 2, tiene los siguientes filtros

$$h(n) = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right] \quad (6.2)$$

para el filtro de paso bajo y :

$$g(n) = \left[\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right] \quad (6.3)$$

para el filtro de paso alto.

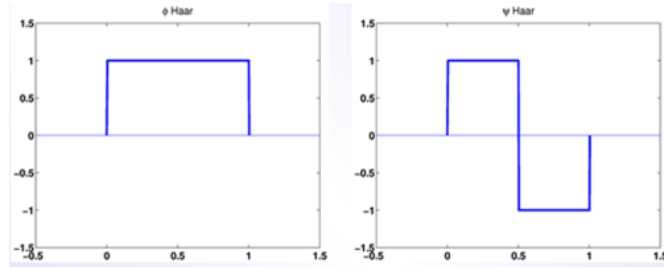


Figura 6.3 Función escala ϕ y función wavelet ψ de Haar.

11

Tiene su función escala simétrica, pero su función wavelet no es simétrica, son ortogonales y de soporte compacto.

Se puede ver que si una función es aproximadamente constante sobre el soporte de wavelet Haar, entonces los detalles $d_{m,n}$ serán aproximadamente cero.

El árbol de descomposición del wavelet de Haar y el nivel aproximado de frecuencias en que se encuentra cada nivel de descomposición es el siguiente:

Luego por cada frame de la señal de habla se hizo una descomposición con las wavelets de Haar.

6.1.2 Wavelet de Daubechies

Los wavelet de Daubechies tienen un soporte ligeramente mayor que el de las wavelets de Haar, pero constituyen una herramienta poderosa para el procesamiento digital de las señales.

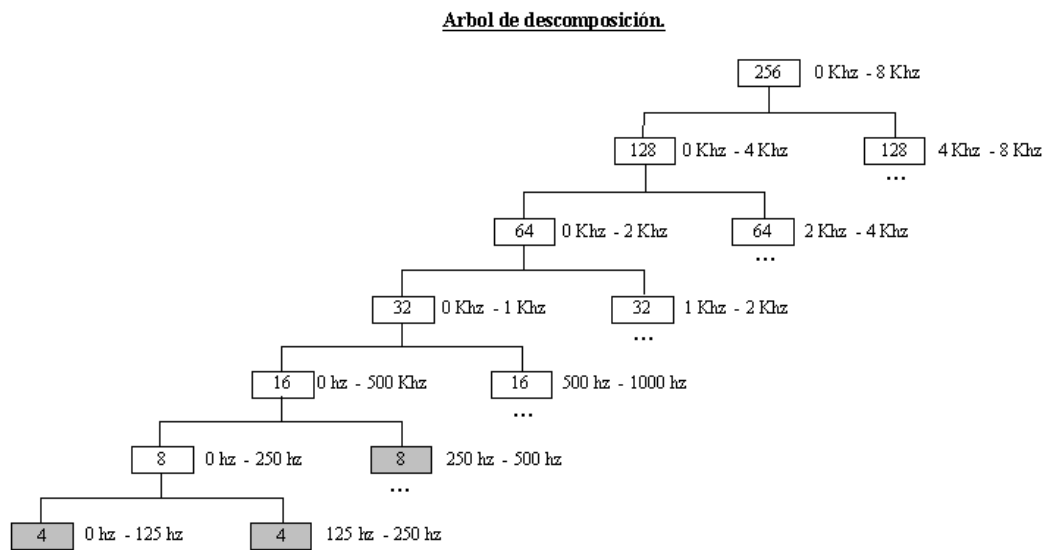


Figura 6.4 Árbol de descomposición para las wavelets de Haar

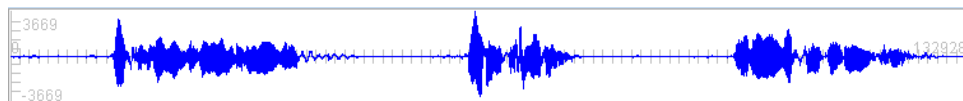


Figura 6.5 Señal de Habla correspondiente a los dígitos 10001-90210-01803.

Existen muchas clases de Wavelets Daubechies, en esta tesis usaremos las wavelets Daubechies 4 y las wavelets Daubechies 6.

Los filtros para las wavelets de Daubechies 4 son:

$$h(n) = \left[\frac{1 + \sqrt{3}}{4\sqrt{2}}, \frac{3 + 3\sqrt{3}}{4\sqrt{2}}, \frac{3 - 3\sqrt{3}}{4\sqrt{2}}, \frac{1 - \sqrt{3}}{4\sqrt{2}} \right] \quad (6.4)$$

para el filtro de paso bajo y :

$$g(n) = \left[\frac{1 - \sqrt{3}}{4\sqrt{2}}, \frac{3\sqrt{3} - 3}{4\sqrt{2}}, \frac{3 + 3\sqrt{3}}{4\sqrt{2}}, \frac{-1 - \sqrt{3}}{4\sqrt{2}} \right] \quad (6.5)$$

para el filtro de paso alto.

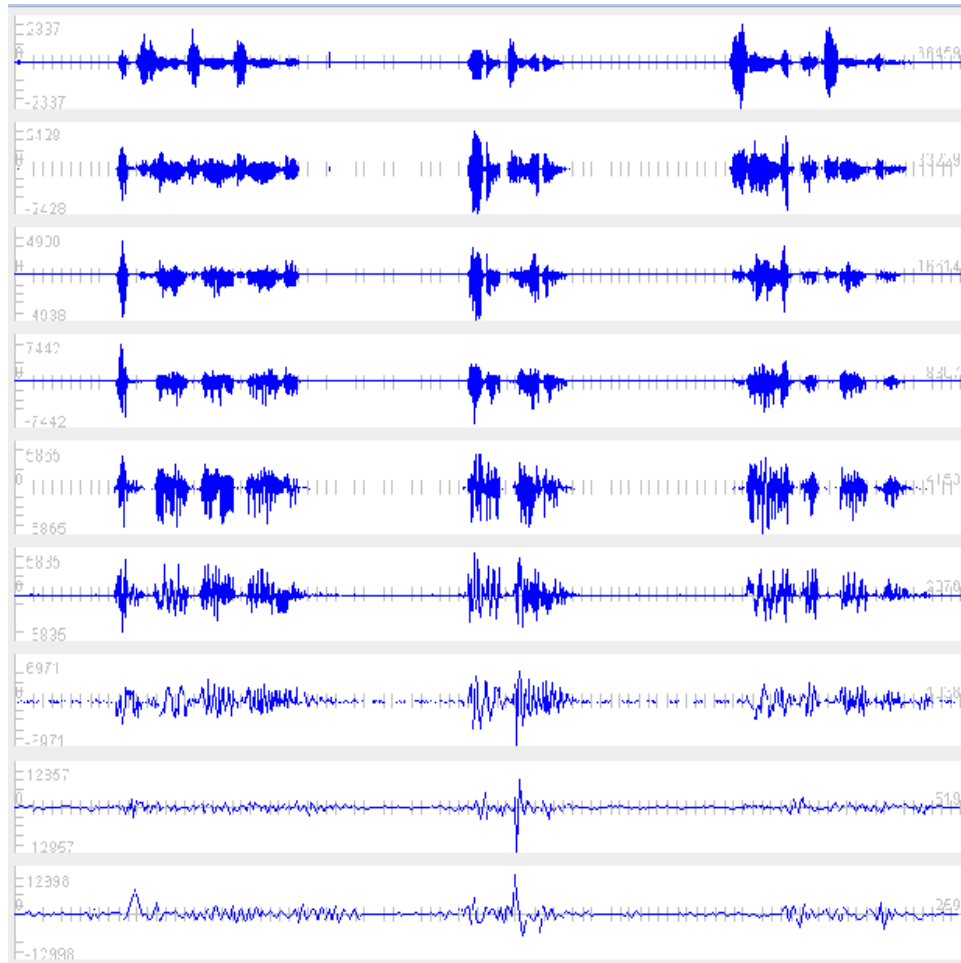


Figura 6.6 Descomposición de la señal de habla correspondiente a la secuencia de dígitos 10001-90210-01803, mediante las wavelets de Haar.

Los filtros para las wavelets de Daubechies 6 son

$$h(n) = [0.3326, 0.8068, 0.4598, -0.1350, -0.0854, 0.0352] \quad (6.6)$$

para el filtro de paso bajo y :

$$g(n) = [0.0352, 0.0854, -0.1350, -0.4598, 0.8068, -0.3326] \quad (6.7)$$

para el filtro de paso alto.

Algunas propiedades importantes de este tipo de wavelets son las siguientes: la energía de sus filtros h es 1, su suma es $\sqrt{2}$, y la suma de los filtros g es cero.

Los valores de los filtros son encontrados como sigue:

Si tenemos un orden J del wavelet de la familia de daubechies, entonces:

$$h_1, \dots, h_J \quad (6.8)$$

Es el tamaño del filtro y además satisfacen las siguientes propiedades:

$$h_1^2 + \dots + h_J^2 = 1 \quad (6.9)$$

$$h_1 + \dots + h_J = \sqrt{2} \quad (6.10)$$

Además para los filtros wavelets g :

$$g_1 = h_J, \quad g_2 = -h_{J-1}, \dots, g_{J-1} = h_2, \quad g_J = -h_1 \quad (6.11)$$

Estos filtros wavelets satisfacen la siguiente propiedad

$$0^L g_1 + 1^L g_2 + \dots + (J-1)^L g_J = 0 \quad (6.12)$$

Donde L puede tomar los valores $[0, J/2 - 1]$ y esto se interpreta como si para una función que es aproximadamente igual a un polinomio de grado menor que $J/2$, sobre el soporte de un wavelet Daubechies J , entonces el valor de los coeficientes en ese nivel son aproximadamente cero. Solucionando matemáticamente esas ecuaciones es como se encuentran los filtros h y g . No existe una función explícita para este tipo de wavelet

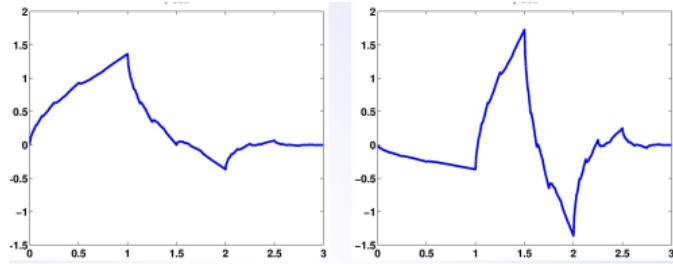


Figura 6.7 Función escala ϕ y función wavelet ψ Daubechies..

En conclusión estos wavelets son asimétricos, de soporte compacto, ortogonales. El árbol de descomposición del wavelet Daubechies 4 es el mismo del wavelet Haar, ver figura 6.4, a continuación el árbol de descomposición para las wavelets de Daubechies 6 y el nivel aproximado de frecuencias por frame en que se encuentra cada nivel de descomposición:

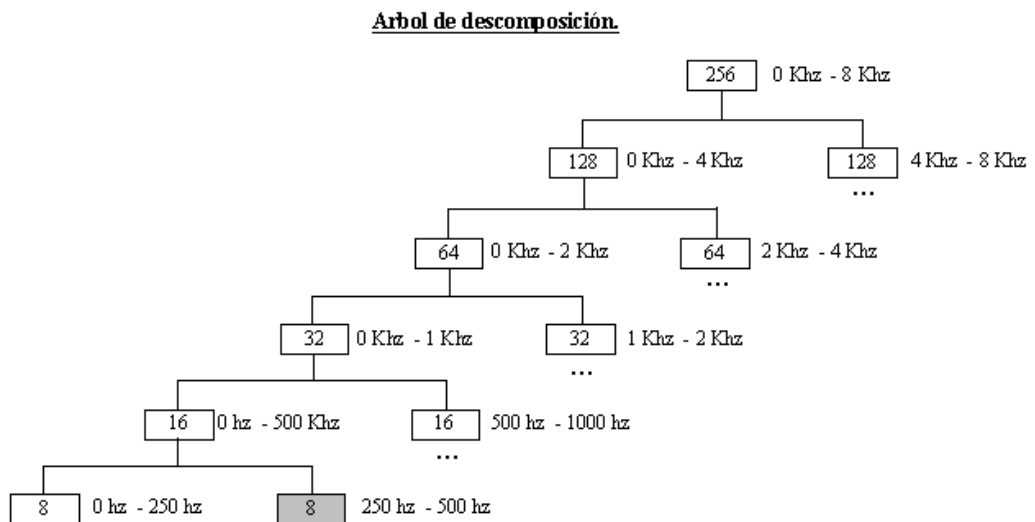


Figura 6.8 Árbol de descomposición para las wavelets de Daubechies 6

Luego por cada frame se hizo una descomposición wavelet

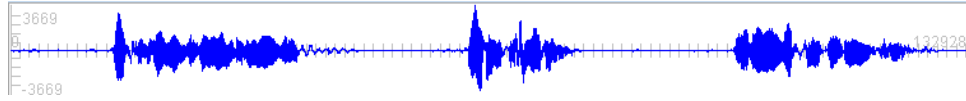


Figura 6.9 Señal de habla correspondiente a los dígitos 10001-90210-01803.

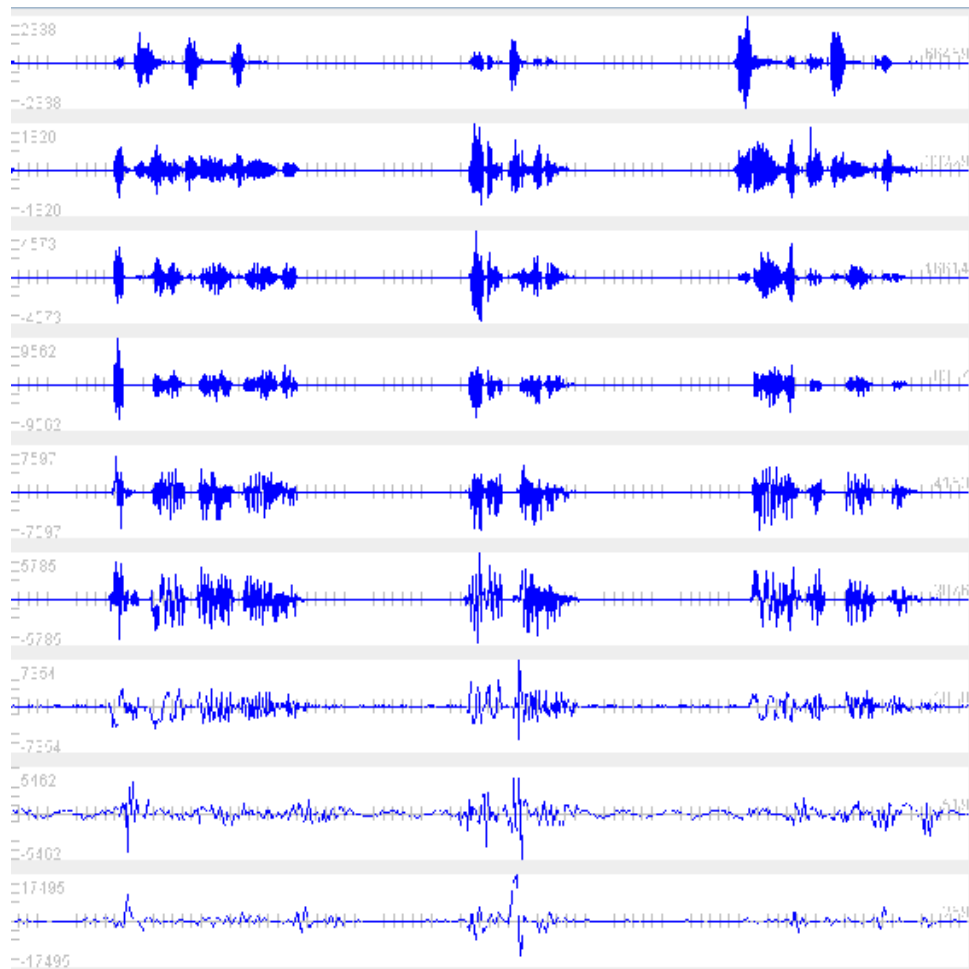


Figura 6.10 Descomposición de la señal de habla correspondiente a la secuencia de dígitos 10001-90210-01803, mediante las wavelets de Daubechies 4

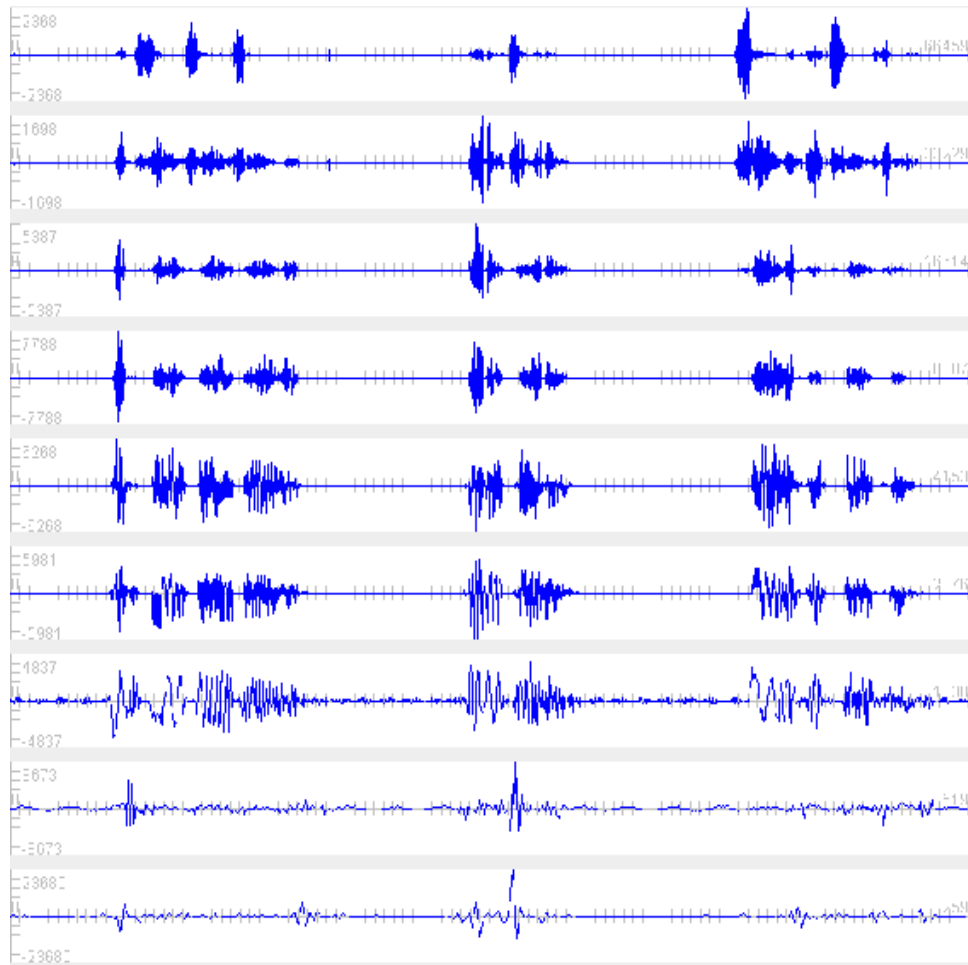


Figura 6.11 Descomposición de la señal de habla correspondiente a la secuencia de dígitos 10001-90210-01803, mediante las wavelets de Daubechies 6

6.1.3 Wavelet Coiflets

Usamos las wavelets Coiflet 6 cuyos filtros están dados por:

$$h_1 = \frac{1 - \sqrt{7}}{16\sqrt{2}}, \quad h_2 = \frac{5 + \sqrt{7}}{16\sqrt{2}}, \quad h_3 = \frac{14 + 2\sqrt{7}}{16\sqrt{2}},$$

$$h_4 = \frac{14 - 2\sqrt{7}}{16\sqrt{2}}, \quad h_5 = \frac{1 - \sqrt{7}}{16\sqrt{2}}, \quad h_6 = \frac{-3 + \sqrt{7}}{16\sqrt{2}}$$

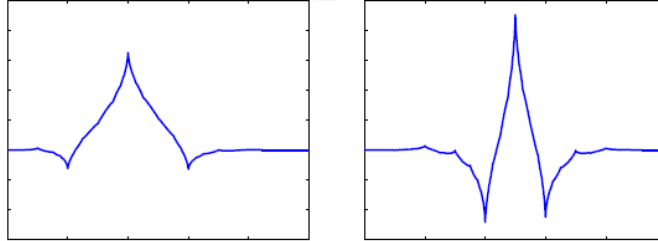


Figura 6.12 Función escala ϕ y función wavelet ψ de Coiflets.

Las wavelets de coiflet satisfacen las siguientes propiedades:

$$h_1^2 + \dots + h_J^2 = 1 \quad (6.13)$$

$$h_1 + \dots + h_J = \sqrt{2} \quad (6.14)$$

Las demás propiedades son muy similares a las de los wavelet de Daubechies, la diferencia está en que los filtros h deben cumplir lo siguiente:

$$-2h_1 - 1h_2 + 0h_3 + 1h_4 + 2h_5 + 3h_6 = 0 \quad (6.15)$$

$$(-2)^2 h_1 + (-1)^2 h_2 + 0^2 h_3 + 1^2 h_4 + 2^2 h_5 + 3^2 h_6 = 0 \quad (6.16)$$

Estas ecuaciones indican que los valores de los coeficientes de aproximación, son promedios de sucesivos valores en la señal, estos wavelets son mucho más simétricos que las wavelets de Daubechies

El árbol de descomposición es el mismo que el de Daubechies 6, ver figura 6.8.

luego por cada frame se hizo una descomposición wavelet

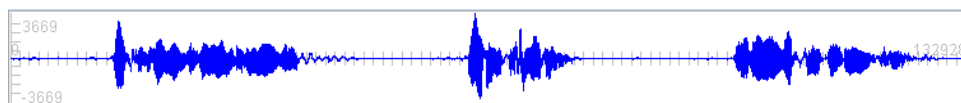


Figura 6.13 Señal de habla correspondiente a los dígitos 10001-90210-01803.

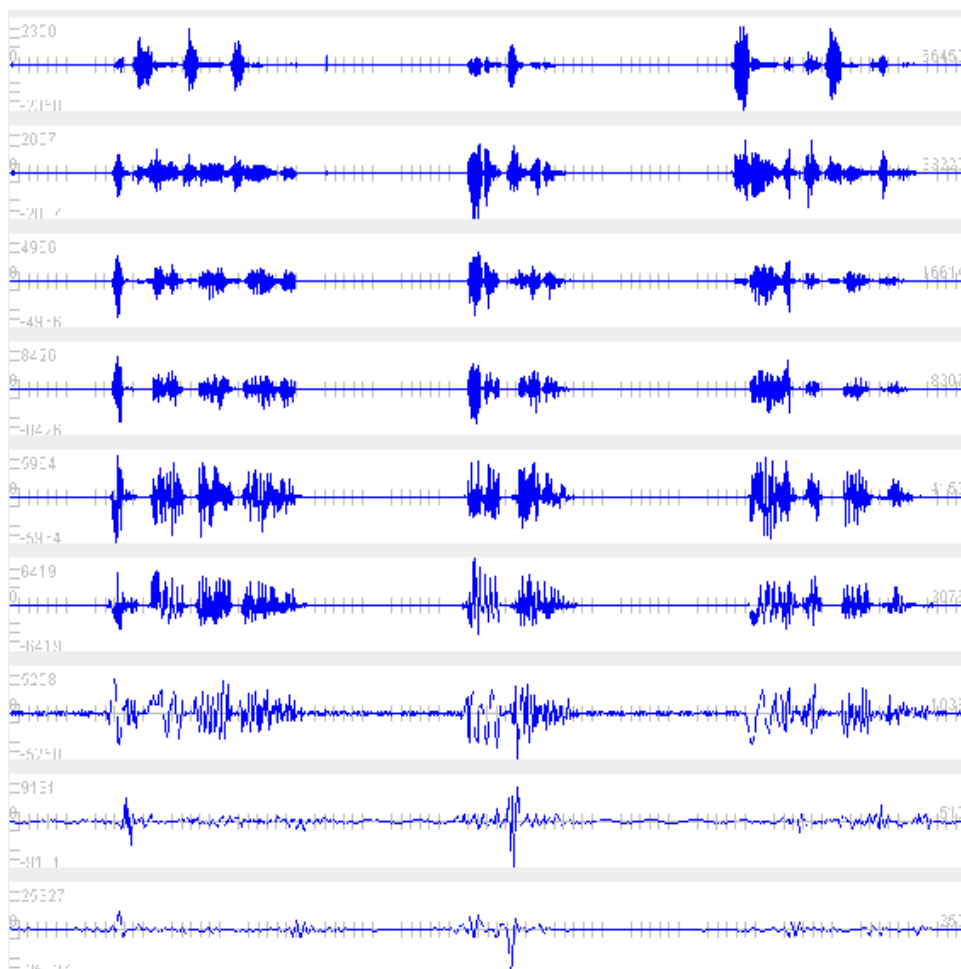


Figura 6.14 Descomposición de la señal de habla correspondiente a la secuencia de dígitos 10001-90210-01803. mediante las wavelets Coiflet 6

6.1.4 Obtención de las Características

Una vez computados los valores de los coeficientes para cada nivel de descomposición se procedió a calcular la energía de cada nivel con la finalidad de comprender el aporte del nivel en el tiempo, la energía de cada nivel se calculó mediante la siguiente expresión:

$$E_i = \frac{\sum_{j=1}^N (W_i^p f(j))^2}{N_i} \quad (6.17)$$

Y finalmente se aplica una transformada discreta del coseno al logaritmo de las energías para cada bando de frecuencia, estos valores son los que constituyeron los vectores de características de la señal.

$$F(i) = \sum_{n=1}^N \log E_n \cos\left(\frac{i\left(\frac{n-1}{2}\right)}{N}\right) \quad (6.18)$$

La complejidad de todo este algoritmo es $O(n)$ ver sección 4.9

6.2 Modelo propuesto para la Extracción de Características basadas en las wavelets packet Perceptual con Daubechies 4

En este modelo que proponemos hacemos uso de los wavelet Packet, pero no nos interesa toda la descomposición del wavelet packet, solo deseamos obtener los coeficientes que están en determinado nivel de resolución, cuyos componentes de frecuencia son aproximadamente iguales a la escala Mel, proponemos 2 modelos, el primero al

igual que la técnica de los Coeficientes Cepstrales en Frecuencia Mel, se basa en el hecho de calcular k valores en la escala mel por medio de filtros triangulares, que para la presente tesis se tomó el valor de $k = 13$, nosotros de igual manera hicimos un análisis multiresolución con los wavelets packets pero teniendo cuidado de construir el árbol de descomposición de tal manera que contenga aproximadamente las frecuencias igualmente espaciadas en escala Mel.

Para este caso utilizamos un tamaño de frame igual a $24mS$ y un tamaño de paso igual que el caso anterior de $16mS$.

Análisis con Wavelet Packet paso (202.45) | MEL

Equivalencias escala MEL a Frecuencias

	MEL	Frecuencias
1	0	0
2	202.10	135
3	404.21	300
4	606.31	500
5	808.42	735
6	1010.53	1030
7	1212.63	1375
8	1414.73	1750
9	1616.85	2235
10	1818.95	2830
11	2021.06	3500
12	2223.16	4333
13	2425.27	5310
14	2627.38	6500
	2829.48	8000

Figura 6.15 Equivalencias de Mel a Frecuencias

Se obtuvieron 14 espacios de resolución con frecuencias aproximadas a la escala Mel. El Proceso de descomposición fué el siguiente:

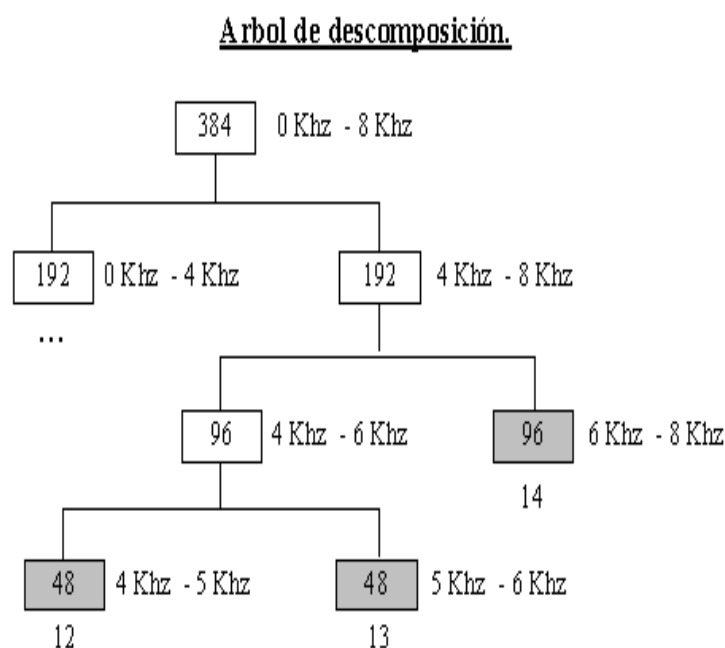


Figura 6.16 Arbol de descomposición espacios de resolución 12,13 y 14

6.2.1 Obtención de las Características para las Wavelets Packet

Una vez computados los valores de los coeficientes para cada nivel de descomposición se procedió a calcular la energía de cada nivel con la finalidad de comprender el aporte del nivel en el tiempo, la energía de cada nivel se calculó mediante la siguiente expresión:

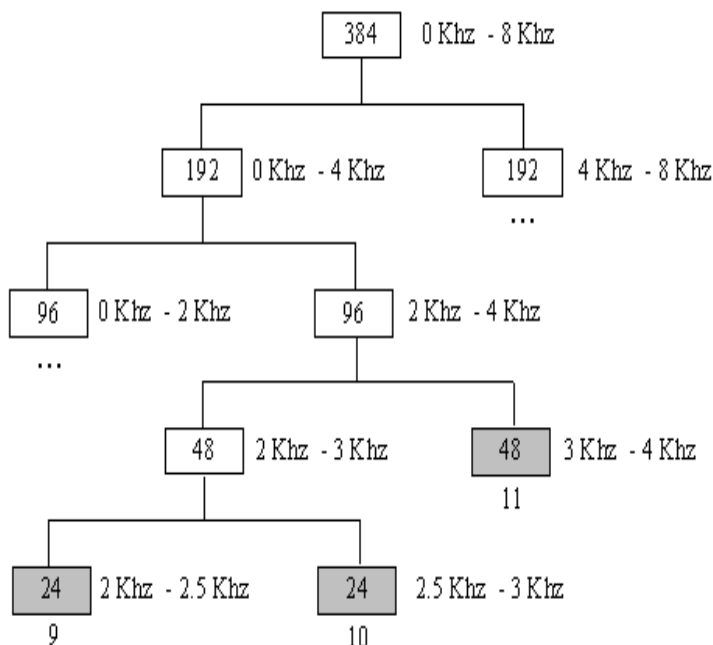
Arbol de descomposición.

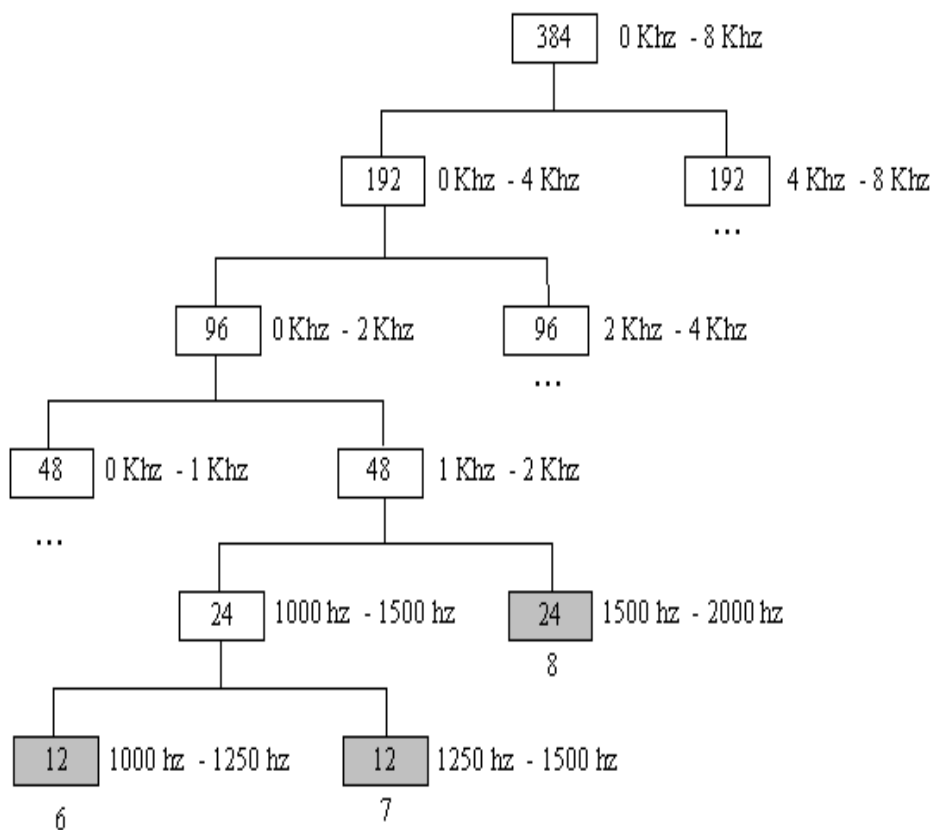
Figura 6.17 Arbol de descomposición espacios de resolución 9,10 y 11

$$E_i = \frac{\sum_{j=1}^N (W_i^p f(j))^2}{N_i} \quad (6.19)$$

Y finalmente se aplica una transformada discreta del coseno al logaritmo de las energías para cada bando de frecuencia, estos valores son los que constituirán los vectores de características de la señal

$$F(i) = \sum_{n=1}^N \log E_n \cos\left(\frac{i\left(\frac{n-1}{2}\right)}{N}\right) \quad (6.20)$$

La complejidad de todo este algoritmo es $O(n \log n)$. Similarmente se implementó

Arbol de descomposición.**Figura 6.18** Arbol de descomposición espacios de resolución 6,7 y 8

un algoritmo basado ya no en 13 espacios si no en 23 espacios, para la obtención de los valores de características no se tomaron los coeficientes de mas baja resolución pues esta resolución es muy afectada por las características de los canales de trasmisión. Las wavelets empleados para generar las wavelets packets fueron las Wavelets de Haar (Wavelet Packet Walsh), Daubechies 4 y Daubechies 6.

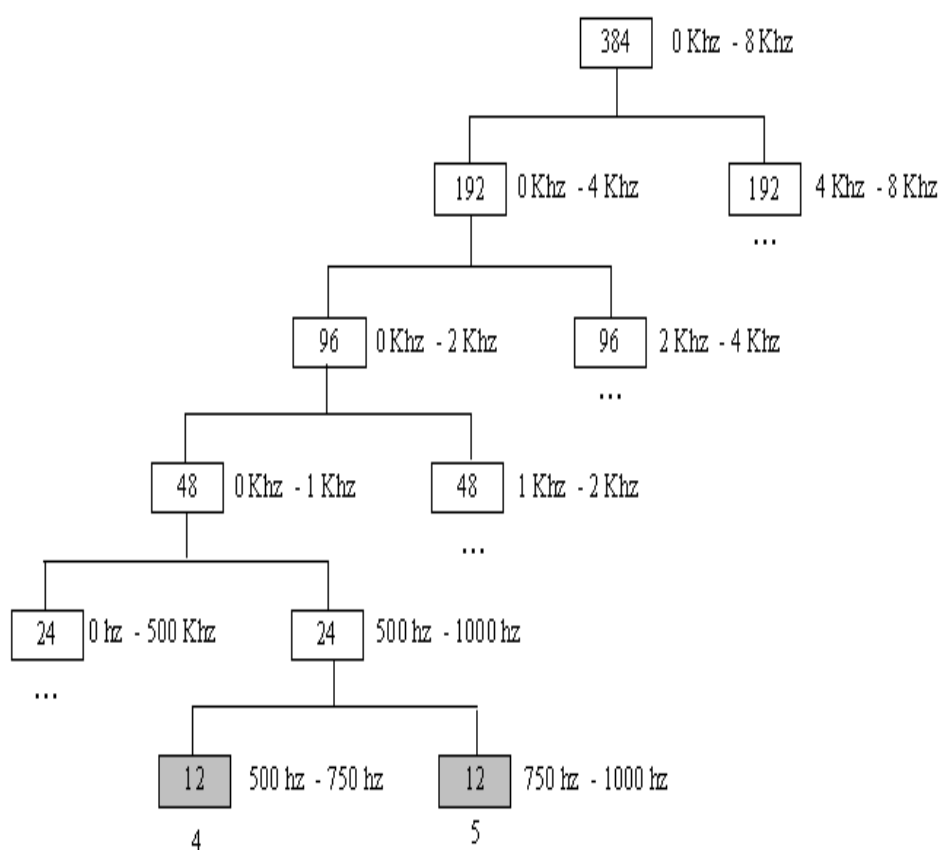
Arbol de descomposición.

Figura 6.19 Arbol de descomposición espacios de resolución 4 y 5

Arbol de descomposición.

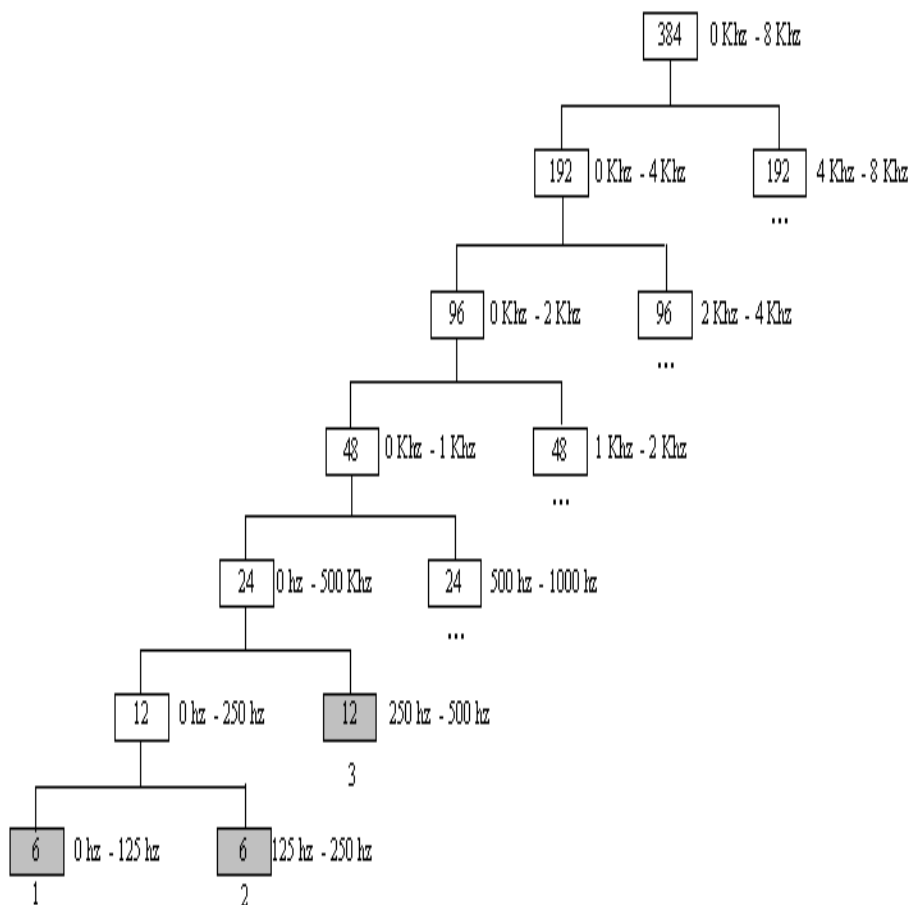


Figura 6.20 Arbol de descomposición espacios de resolución 1, 2 y 3

Parte II

Métodos

Capítulo 7

Métodos y Técnicas

7.1 Enfoque

La presente tesis ha sido desarrollada siguiendo un enfoque cuantitativo pues este utiliza técnicas como contar, medir y usar un razonamiento abstracto para interpretar los resultados, más no apreciaciones subjetivas [Barrantes and Rodrigo, 2001].

7.2 Hipótesis

La hipótesis planteada fué la siguiente:

El procesamiento digital de la señal del habla con wavelets nos proporciona una alternativa frente a un análisis de la señal mediante la Transformada de Fourier, mejorando el análisis del espectro, y disminuyendo la complejidad computacional en el análisis de la señal, como paso previo para la obtención de vectores de características en el reconocimiento automático del habla por parte de un ordenador.

7.3 Tipo de Investigación

Se aplicó el modelo experimental , pues el presente trabajo tiene características fundamentales de dicho modelo.

7.4 Universo y Muestra

El universo es el conjunto de palabras siguiente: *abrir, cerrar, coger, cuatro, dos, eliminar, error, hola, izquierda, pez, salir, terminar, tres, tres(repetición), uno.*

La muestra constituyen 900 palabras grabadas de 60 personas de sexo masculino con una edad promedio de 23 años, la muestra no es significativa, pero por tratarse de un piloto, es suficiente para realizar los experimentos [Pérez, 2000].

7.5 Instrumentos

Para la recolección de datos se usó:

- Un micrófono marca Shure modelo C606 semiprofesional.
- Una computadora portátil marca Dell centrino doble nucleo 1.73 G.
- Una tarjeta de sonido sigmatel stac 9245 C-Major HD Audio.

Para las pruebas se desarrolló un software el cual fué denominado LORITO.

7.6 Procedimiento

El desarrollo de la siguiente investigación usó el método científico, a continuación los pasos en el desarrollo de la investigación

1. Revisión bibliográfica
2. Delimitación del problema
3. Elaboración de la realidad Problemática
4. Formulación de la Hipótesis

5. Elaboración del software de pruebas
6. Recolección de datos
7. Pruebas
8. Documentación del Informe

7.7 Métodos y procedimientos para la recolección de datos

Para recolectar los datos se hicieron grabaciones a diferentes personas en diferentes salones de la Universidad Nacional de Trujillo, para esto se contó con un micrófono y una computadora portátil, el procedimiento empleado fué el darle a cada persona una lista de 25 palabras, las cuales al ser leídas , eran a la vez grabadas.

El nivel de ruido de la grabación fué moderado.

7.8 Análisis estadísticos de los datos

Para efectuar el análisis estadísticos de los datos se hizo la Prueba ji-cuadrado de Mc Nemar para datos correlacionados.

Parte III

Resultados y Análisis

Capítulo 8

Resultados

En este capítulo mostramos los resultados obtenidos en la presente tesis.

El tipo de prueba que se realizó fué independiente del hablante, tomando como plantillas para el reconocedor las voces de 5 personas, es decir 125 palabras de entrenamiento.

Para el entrenamiento se utilizaron 25 palabras por persona, las palabras fueron: *abajo, abrir, adiós, arriba, caminar, cancelar, cerrar, coger, cuatro, derecha, detener, dos, eliminar, error, guardar, hola, iniciar, izquierda, pez, salir, terminar, test, tres, tres(repetición), uno*. Para las pruenas se utilizaron 15 palabras por persona, las palabras fueron: *abrir, cerrar, coger, cuatro, dos, eliminar, error, hola, izquierda, pez, salir, terminar, tres, tres(repetición), uno*.

TABLA 8.1

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL MÉTODO DE FOURIER Y EL MÉTODO PROPUESTO BASADO EN LAS WAVELET HAAR.

TRUJILLO 2006.

MÉTODO WAVELET HAAR	MÉTODO DE FOURIER		TOTAL
	Palabras identificadas	Palabras no identificadas	
- Palabras identificadas	262	13	275
- Palabras no identificadas	172	453	625
TOTAL	434	466	900

χ^2_{MCM} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{MCM} = 136.65$	$p < 0.01$
-------------------------	------------

$$\text{Eficiencia relativa de aciertos} = 275 / 434 \times 100$$

Figura 8.1 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets de Haar. Trujillo 2006.

TABLA 8.2

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL METODO DE FOURIER Y EL MÉTODO PROPUESTO BASADO EN LAS WAVELETS DE DAUBECHIES 4.

TRUJILLO 2006.

MÉTODO WAVELET DAUB 4		MÉTODO DE FOURIER		TOTAL
		Palabras identificadas	Palabras no identificadas	
-	Palabras identificadas	430	6	436
-	Palabras no identificadas	4	460	464
TOTAL		434	466	900

χ^2_{MN} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{MN} = 0.40$	$p > 0.05$
----------------------	------------

Eficiencia relativa de aciertos = $436 / 434 \times 100$

ER-aciertos = 100.46

Figura 8.2 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets de Daubechies 4. Trujillo 2006.

TABLA 8.3

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL METODO DE FOURIER Y EL MÉTODO PROPUESTO BASADO EN LAS WAVELETS DE DAUBECHIES 6 .

TRUJILLO 2006.

MÉTODO WAVELET DAUB 6	MÉTODO DE FOURIER		TOTAL
	Palabras identificadas	Palabras no identificadas	
- Palabras identificadas	430	93	523
- Palabras no identificadas	4	373	377
TOTAL	434	466	900

χ^2_{MIN} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{MIN} = 81.66$	$p < 0.01$
------------------------	------------

Eficiencia relativa de aciertos = $523 / 434 \times 100$

ER-aciertos = 120.51

Figura 8.3 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets de Daubechies 6. Trujillo 2006.

TABLA 8.4

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL MÉTODO DE FOURIER Y EL MÉTODO PROPUESTO BASADO EN LAS LAS WAVELETS DE COIFLETS6.

TRUJILLO 2006.

MÉTODO WAVELET PACKET COIFLETS 6		MÉTODO DE FOURIER		TOTAL
		Palabras identificadas	Palabras no identificadas	
-	Palabras identificadas	418	51	469
-	Palabras no identificadas	16	415	431
TOTAL		434	466	900

χ^2_{Mn} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{Mn} = 18.28$	$p < 0.01$
-----------------------	------------

$$\text{Eficiencia relativa de aciertos} = 469 / 434 \times 100$$

$$\text{ER-aciertos} = 108.06$$

Figura 8.4 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets de Coiflet 6. Trujillo 2006.

TABLA 8.5

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL METODO DE FOURIER Y EL MÉTODO PROPUESTO BASADO EN LAS WAVELETS PACKET WALSH.

TRUJILLO 2006.

MÉTODO WAVELET PACKET WALSH	MÉTODO DE FOURIER		TOTAL
	Palabras identificadas	Palabras no identificadas	
- Palabras identificadas	390	89	479
- Palabras no identificadas	44	377	421
TOTAL	434	466	900

χ^2_{Mn} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{Mc} = 15.23$	$p < 0.01$
-----------------------	------------

Eficiencia relativa de aciertos = $479 / 434 \times 100$

ER-aciertos = 110.37

Figura 8.5 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets Packet Walsh. Trujillo 2006.

TABLA 8.6

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL METODO DE FOURIER Y EL METODO PROPUESTO BASADO EN LAS WAVELETS PACKET DAUBECHIES 4.

TRUJILLO 2006.

MÉTODO WAVELET PACKET DAUB 4	MÉTODO DE FOURIER		TOTAL
	Palabras identificadas	Palabras no identificadas	
- Palabras identificadas	428	170	598
- Palabras no identificadas	6	296	302
TOTAL	434	466	900

χ^2_{Mc} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{Mc} = 152.82$	$p < 0.01$
------------------------	------------

Eficiencia relativa de aciertos = $498 / 434 \times 100$

ER-aciertos = 137.79

Figura 8.6 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets Packet Daubechies 4. Trujillo 2006.

TABLA 8.7

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL METODO DE FOURIER Y EL MÉTODO PROPUESTO BASADO EN LAS WAVELETS PACKET DAUBECHIES 6.

TRUJILLO 2006.

MÉTODO WAVELET PACKET DAUB6		MÉTODO DE FOURIER		TOTAL
		Palabras identificadas	Palabras no identificadas	
-	Palabras identificadas	425	220	645
-	Palabras no identificadas	9	246	255
TOTAL		434	466	900

χ^2_{MNN} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{MNN} = 194.41$	$p < 0.01$
-------------------------	------------

Eficiencia relativa de aciertos = $645 / 434 \times 100$

ER-aciertos = 148.62

Figura 8.7 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets packet Daubechies 6. Trujillo 2006.

TABLA 8.8

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL METODO DE FOURIER Y EL METODO PROPUESTO BASADO EN LAS WAVELETS PACKET PERCEPTUAL CON DAUBECHIES 4.
TRUJILLO 2006.

MÉTODO WAVELET PACKET PERCEPTUAL	MÉTODO DE FOURIER		TOTAL
	Palabras identificadas	Palabras no identificadas	
- Palabras identificadas	420	183	603
- Palabras no identificadas	14	283	297
TOTAL	434	466	900

χ^2_{MIN} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{Mc} = 144.98$	$p < 0.01$
------------------------	------------

$$\text{Eficiencia relativa de aciertos} = 603 / 434 \times 100$$

$$\text{ER-aciertos} = 138.94$$

Figura 8.8 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método de Fourier y el método propuesto basado en las wavelets packet Perceptual con Daubechies 4. Trujillo 2006.

TABLA 8.9

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL METODO MFCC Y EL PROPUESTO BASADO EN LAS WAVELETS HAAR.

TRUJILLO 2006.

MÉTODO WAVELET HAAR	MÉTODO MFCC		TOTAL
	Palabras identificadas	Palabras no identificadas	
- Palabras identificadas	275	0	275
- Palabras no identificadas	456	169	625
TOTAL	731	169	900

χ^2_{MN} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{MN} = 456.0$	$p < 0.01$
-----------------------	------------

Eficiencia relativa de aciertos = $275 / 731 \times 100$

Figura 8.9 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets de Haar. Trujillo 2006.

TABLA 8.10

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL METODO MFCC Y EL PROPUESTO BASADO EN LAS WAVELETS DAUBECHIES 4 TRUJILLO 2006.

MÉTODO WAVELET DAUB 4	MÉTODO MFCC		TOTAL
	Palabras identificadas	Palabras no identificadas	
- Palabras identificadas	433	3	436
- Palabras no identificadas	298	166	464
TOTAL	731	169	900

χ^2_{McN} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{McN} = 289.12$	$p < 0.01$
-------------------------	------------

Eficiencia relativa de aciertos = $436 / 731 \times 100$

ER-aciertos = 59.64

Figura 8.10 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets de Daubechies 4. Trujillo 2006.

TABLA 8.11

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL MÉTODO MFCC Y EL PROPUESTO BASADO EN LAS WAVELETS DAUBECHIES 6. TRUJILLO 2006.

MÉTODO WAVELET DAUB 6	MÉTODO MFCC		TOTAL
	Palabras identificadas	Palabras no identificadas	
- Palabras identificadas	518	5	523
- Palabras no identificadas	213	164	377
TOTAL	731	169	900

χ^2_{McN} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{McN} = 198.46$	$p < 0.01$
-------------------------	------------

Eficiencia relativa de aciertos = $523 / 731 \times 100$

ER-aciertos = 71.55

Figura 8.11 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets de Daubechies 6. Trujillo 2006.

TABLA 8.12

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL METODO MFCC Y EL PROPUESTO BASADO EN LAS WAVELETS COIFLETS6.

TRUJILLO 2006.

MÉTODO WAVELET PACKET COIFLETS 6		MÉTODO MFCC		TOTAL
		Palabras identificadas	Palabras no identificadas	
-	Palabras identificadas	460	9	469
-	Palabras no identificadas	271	160	431
TOTAL		731	169	900

χ^2_{Mc} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{Mc} = 245.16$	$p < 0.01$
------------------------	------------

Eficiencia relativa de aciertos = $469 / 731 \times 100$

ER-aciertos = 64.16

Figura 8.12 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets de Coiflet 6. Trujillo 2006.

TABLA 8.13

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL MÉTODO MFCC Y EL PROPUESTO BASADO EN LAS WAVELETS PACKET WALSH.

TRUJILLO 2006.

MÉTODO WAVELET PACKET WALSH	MÉTODO MFCC		TOTAL
	Palabras identificadas	Palabras no identificadas	
- Palabras identificadas	479	0	479
- Palabras no identificadas	252	169	421
TOTAL	731	169	900

χ^2_{MIN} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{Mc} = 252.0$	$p < 0.01$
-----------------------	------------

Eficiencia relativa de aciertos = $479 / 731 \times 100$

ER-aciertos = 65.53

Figura 8.13 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets Packet Walsh. Trujillo 2006.

TABLA 8.14

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL MÉTODO MFCC Y EL PROPUESTO BASADO EN LAS WAVELETS PACKET DAUBECHIES 4.

TRUJILLO 2006.

MÉTODO WAVELET PACKET DAUB 4		MÉTODO MFCC		TOTAL
		Palabras identificadas	Palabras no identificadas	
-	Palabras identificadas	597	1	598
-	Palabras no identificadas	134	168	302
TOTAL		731	169	900

χ^2_{MN} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{cr} = 131.03$	$p < 0.01$
------------------------	------------

Eficiencia relativa de aciertos = $598 / 731 \times 100$

ER-aciertos = 81.81

Figura 8.14 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets Packet Daubechies 4. Trujillo 2006.

TABLA 8.15

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL MÉTODO MFCC Y EL PROPUESTO BASADO EN LAS WAVELETS PACKET DAUBECHIES 6.

TRUJILLO 2006.

MÉTODO WAVELET PACKET DAUB6		MÉTODO MFCC		TOTAL
		Palabras identificadas	Palabras no identificadas	
-	Palabras identificadas	640	5	645
-	Palabras no identificadas	91	164	255
TOTAL		731	169	900

χ^2_{Mc} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{Mc} = 77.04$	$p < 0.01$
-----------------------	------------

$$\text{Eficiencia relativa de aciertos} = 645 / 731 \times 100$$

$$\text{ER-aciertos} = 88.24$$

Figura 8.15 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets packet Daubechies 6. Trujillo 2006.

TABLA 8.16

DISTRIBUCION DE 900 PALABRAS PRONUNCIADAS POR 60 PERSONAS SEGUN SU IDENTIFICACION Y NO IDENTIFICACION POR LA COMPUTADORA CON EL METODO MFCC Y EL PROPUESTO BASADO EN LAS WAVELETS PACKET PERCEPTUAL CON DAUBECHIES 4.
TRUJILLO 2006.

MÉTODO WAVELET PACKET PERCEPTUAL	MÉTODO MFCC		TOTAL
	Palabras identificadas	Palabras no identificadas	
- Palabras identificadas	600	3	603
- Palabras no identificadas	131	164	297
TOTAL	731	169	900

χ^2_{Mn} : Prueba ji-cuadrado de Mc Nemar para datos correlacionados

$\chi^2_{Mc} = 122.27$	$p < 0.01$
------------------------	------------

$$\text{Eficiencia relativa de aciertos} = 603 / 731 \times 100$$

$$\text{ER-aciertos} = 82.49$$

Figura 8.16 Distribución de 900 palabras pronunciadas por 60 personas según su identificación y no identificación por la computadora con el método MFCC y el método propuesto basado en las wavelets packet Perceptual con Daubechies 4. Trujillo 2006.

A continuación se muestran los resultados obtenidos por los métodos implementados y la tasa de reconocimiento de palabras por método.

<i>Método</i>	<i>Tasa de aceptación</i>	<i>Error</i>
Coeeficientes Cepstrales en Escala Mel	81.61%	18.39%
Wavelet Haar	31.03%	68.97%
Wavelet Daubechies 4	48.74%	51.26%
Wavelet Daubechies 6	58.39 %	41.61%
Wavelet Coiflets 6	52.18 %	47.82%
Wavelet Packet Perceptuales Walsh	53.33 %	46.67 %
Wavelet Packet Perceptuales Daubechies 4	66.67 %	33.33%
Wavelet Packet Perceptuales Daubechies 6	71.49 %	28.51%
Wavelet Packet Perceptuales Daubechies 4 (22)	67.13 %	32.87%

Tabla 8.1 Datos obtenidos utilizando la técnica de DTW como reconocedor, con distancia Euclidiana y con Slope Constrain P=0. Se observa la mejor performance en las Wavelet Packet Perceptuales Daubechies 6, y la mas pobre en las Wavelet Haar

<i>Método</i>	<i>Tasa aceptación</i>	<i>Error</i>
Coeficientes Cepstrales en Escala Mel	82.53%	17.47%
Wavelet Haar	32.83%	67.17%
Wavelet Daubechies 4	50.01%	49.99%
Wavelet Daubechies 6	58.47 %	41.53%
Wavelet Coiflets 6	53.46%	46.54%
Wavelet Packet Perceptuales Walsh	53.74 %	46.26%
Wavelet Packet Perceptuales Daubechies 4	68.43 %	31.57%
Wavelet Packet Perceptuales Daubechies 6	73.45 %	26.55%
Wavelet Packet Perceptuales Daubechies 4 (22)	69.25%	30.75%

Tabla 8.2 Datos obtenidos utilizando la técnica de DTW como reconocedor, con distancia Euclidiana y con Slope Constrain P=1. Se observa la mejor performance en las Wavelet Packet Perceptuales Daubechies 6, y la mas pobre en las Wavelet Haar

<i>Método</i>	<i>Tasa aceptación</i>	<i>Error</i>
Coeficientes Cepstrales en Escala Mel	81.32%	18.68%
Wavelet Haar	33.33%	66.67%
Wavelet Daubechies 4	49.56%	50.44%
Wavelet Daubechies 6	59.46 %	46.77%
Wavelet Coiflets 6	55.78 %	44.22%
Wavelet Packet Perceptuales Walsh	55.33 %	44.67 %
Wavelet Packet Perceptuales Daubechies 4	68.45 %	31.55%
Wavelet Packet Perceptuales Daubechies 6	73.46 %	26.54%
Wavelet Packet Perceptuales Daubechies 4 (22)	68.46 %	31.54%

Tabla 8.3 Datos obtenidos utilizando la técnica de DTW como reconocedor, con distancia Chebyshev y con Slope Constrain P=0. Se observa la mejor performance en las Wavelet Packet Perceptuales Daubechies 6, y la mas pobre en las Wavelet Haar

<i>Método</i>	<i>Tasa aceptación</i>	<i>Error</i>
Coefficientes Cepstrales en Escala Mel	85.32%	14.68%
Wavelet Haar	34.47%	65.53%
Wavelet Daubechies 4	51.79%	48.21%
Wavelet Daubechies 6	61.32 %	38.68%
Wavelet Coiflets 6	55.46 %	44.54%
Wavelet Packet Perceptuales Walsh	55.79 %	44.21 %
Wavelet Packet Perceptuales Daubechies 4	69.14 %	30.86%
Wavelet Packet Perceptuales Daubechies 6	74.43 %	25.57%
Wavelet Packet Perceptuales Daubechies 4 (22)	71.21 %	28.79%

Tabla 8.4 Datos obtenidos utilizando la técnica de DTW como reconocedor, con distancia Chebyshev y con Slope Constrain P=1. Se observa la mejor performance en las Wavelet Packet Perceptuales Daubechies 6, y la mas pobre en las Wavelet Haar

	<i>MFCC</i>	<i>W. Haar</i>	<i>W. Db4</i>	<i>W. Db6</i>	<i>W. Coif6</i>	<i>WP Walsh</i>	<i>WP Db4</i>	<i>WP Db6</i>	<i>WP Perc.</i>
<i>Arriba</i>	86%	17%	41%	48%	34%	31%	66%	72%	62
<i>Cerrar</i>	100%	45%	79%	86%	76%	97%	100%	97%	97%
<i>Coger</i>	90%	17%	17%	31%	17%	28%	52%	69%	62%
<i>Cuatro</i>	97%	48%	28%	31%	24%	62%	76%	83%	83%
<i>Dos</i>	86%	28%	38%	48%	48%	59%	69%	69%	72%
<i>Eliminar</i>	72%	24%	34%	52%	48%	24%	48%	76%	45%
<i>Error</i>	97%	24%	66%	93%	83%	93%	97%	97%	93%
<i>Hola</i>	83%	38%	48%	72%	55%	38%	52%	59%	52%
<i>Izquierda</i>	86%	31%	48%	62%	59%	24%	59%	72%	69%
<i>Pez</i>	62%	24%	52%	41%	66%	62%	59%	59%	55%
<i>Salir</i>	86%	66%	79%	79%	72%	76%	76%	79%	76%
<i>Terminar</i>	62%	24%	24%	31%	28%	41%	55%	55%	55%
<i>Tres</i>	83%	24%	52%	59%	45%	41%	66%	62%	66%
<i>Tres</i>	59%	24%	45%	59%	41%	48%	55%	59%	62%
<i>Uno</i>	76%	31%	79%	83%	86%	76%	72%	66%	59%

Tabla 8.5 Tasa de reconocimiento de las palabras por método.

Capítulo 9

Análisis

Para comprobar nuestra hipótesis, en las tablas del capítulo anterior hemos comparado ocho métodos propuestos basados en wavelets, frente a un método basado en el análisis de la señal mediante Fourier (ver 3.5.4), el mejoramiento del análisis de espectro es reflejado en las tasas de reconocimiento de los métodos, y la mejora de la complejidad computacional fué descrita en las secciones :3.3.2, 3.3.5 y 4.9, en donde la complejidad computacional del algoritmo de banco de filtros para las wavelets discretas es $O(n)$, frente a una complejidad computacional de $O(n \log n)$ para la transformada rápida de Fourier. En lo que respecta a las wavelets packets si bien es cierto que una descomposición total tiene una complejidad de $O(n \log n)$, el tiempo de ejecución es mucho menor, en nuestro caso como no hacemos una descomposición total, si no que solamente necesitamos espacios de resolución que tengan frecuencias aproximadas a la frecuencia Mel, el tiempo de ejecución es aún menor.

Esto nos llevó a decir que existe una mejora evidente en cuanto a la complejidad computacional en los algoritmos basados en las wavelets frente a los basados en la Transformada de Fourier.

Para trabajos futuros, también se comparó los métodos basados en las wavelets con una de las técnicas más robustas existentes en la actualidad, esto es con los MFCC,

el análisis de estos resultados son muy importantes pues en base a estos resultados se puede mejorar los métodos propuestos.

A continuación se hace un análisis de los resultados por tabla.

- En la tabla 8.1 se comparan el método de extracción de características basado en Fourier con el método basado en las wavelets de Haar, al aplicar la prueba ji-cuadrado de Mc Nemar para datos correlacionados se se tiene un valor de $p < 0.01$ esto indica que existe una diferencia altamente significativa , es por esto que se rechaza la hipótesis planteada en este caso las wavelets de Haar no constituyen una alternativa frente al método basado en Fourier, con esto se concluye que tiene una pobre resolución de espectro a pesar de tener menor complejidad computacional. La eficiencia relativa de aciertos de la wavelets de Haar frente al método de Fourier es muy baja en este caso.
- En la tabla 8.2 se comparan el método de extracción de características basado en Fourier con el método basado en las wavelets de Daubechies 4, al aplicar la prueba ji-cuadrado de Mc Nemar para datos correlacionados se tiene un valor de $p > 0.05$ esto indica que no existe diferencia significativa , además la eficiencia relativa de aciertos indica que pueden usarse indistintamente cualquiera de los dos métodos, en este caso la hipótesis planteada es verdadera, las wavelets de Daubechies 4 constituyen una alternativa frente al método basado en Fourier, la resolución de espectro es similar pero las wavelets de Daubechies 4 tienen

menor complejidad computacional que la Transformada de Fourier.

- En la tabla 8.3 se comparan el método de extracción de características basado en Fourier con el método basado en las wavelets de Daubechies 6, al aplicar la prueba ji-cuadrado de Mc Nemar para datos correlacionados se tiene un valor de $p < 0.01$ esto indica que existe diferencia altamente significativa , además la eficiencia relativa de aciertos es superior en un 20.51% esto indica la superioridad del método basado en las wavelets Daubechies 6 frente al basado en la Transformada de Fourier, en este caso la hipótesis planteada es verdadera, las wavelets de Daubechies 6 constituyen una alternativa frente al método basado en la Transformada de Fourier, la resolución de espectro es mejor y además las wavelets de Daubechies 6 tienen menor complejidad computacional que la Transformada de Fourier.
- En la tabla 8.4 se comparan el método de extracción de características basado en Fourier con el método basado en las wavelets de Coiflet 6, al aplicar la prueba ji-cuadrado de Mc Nemar para datos correlacionados se tiene un valor de $p < 0.01$ esto indica que existe diferencia altamente significativa , además la eficiencia relativa de aciertos es superior en 8.51% esto indica la superioridad del método basado en las wavelets Coiflets 6 frente al basado en la Transformada de Fourier, en este caso la hipótesis planteada es verdadera, las wavelets de Coiflet

6 constituyen una alternativa frente al método basado en la Transformada de Fourier, la resolución de espectro es mejor y además las wavelets de Coiflets 6 tienen menor complejidad computacional que la Transformada de Fourier.

- En la tabla 8.5 se comparan el método de extracción de características basado en Fourier con el método basado en las wavelets packet Walsh, al aplicar la prueba ji-cuadrado de Mc Nemar para datos correlacionados se tiene un valor de $p < 0.01$ esto indica que existe diferencia altamente significativa , además la eficiencia relativa de aciertos es superior en 10.37% esto indica la superioridad del método basado en las wavelets packet Walsh frente al basado en la Transformada de Fourier, en este caso la hipótesis planteada es verdadera, las wavelets packet walsh constituyen una alternativa frente al método basado en la Transformada de Fourier, la resolución de espectro es mejor, las wavelets packet walsh tienen igual complejidad computacional que la Transformada de Fourier pero el tiempo de ejecución es menor.
- En la tabla 8.6 se comparan el método de extracción de características basado en Fourier con el método basado en las wavelets packet Daubechies 4, al aplicar la prueba ji-cuadrado de Mc Nemar para datos correlacionados se tiene un valor de $p < 0.01$ esto indica que existe diferencia altamente significativa , además la eficiencia relativa de aciertos es superior en 37.79% esto indica la superioridad

del método basado en las wavelets Packet Daubechies 4 frente al basado en la Transformada de Fourier, en este caso la hipótesis planteada es verdadera, las wavelets packet Daubechies 4 constituyen una alternativa frente al método basado en la Transformada de Fourier, la resolución de espectro es mejor, las wavelets packet Daubechies 4 tienen igual complejidad computacional que la Transformada de Fourier pero el tiempo de ejecución es menor.

- En la tabla 8.7 se comparan el método de extracción de características basado en Fourier con el método basado en las wavelets packet Daubechies 6, al aplicar la prueba ji-cuadrado de Mc Nemar para datos correlacionados se tiene un valor de $p < 0.01$ esto indica que existe diferencia altamente significativa, además la eficiencia relativa de aciertos es superior en 48.62% esto indica la superioridad del método basado en las wavelets Packet Daubechies 6 frente al basado en la Transformada de Fourier, en este caso la hipótesis planteada es verdadera, las wavelets packet Daubechies 6 constituyen una alternativa frente al método basado en la Transformada de Fourier, la resolución de espectro es mejor, las wavelets packet Daubechies 6 tienen igual complejidad computacional que la Transformada de Fourier pero el tiempo de ejecución es menor. Este método basado en wavelets packet Daubechies 6 es el que mayor tasas de reconocimientos tiene entre los demás métodos basados en wavelets.

- En la tabla 8.8 se comparan el método de extracción de características basado en Fourier con el método basado en las wavelets packet Perceptual con Daubechies 4, al aplicar la prueba ji-cuadrado de Mc Nemar para datos correlacionados se tiene un valor de $p < 0.01$ esto indica que existe diferencia altamente significativa, además la eficiencia relativa de aciertos es superior en 38.94% esto indica la superioridad del método basado en las wavelets Packet Daubechies 4 frente al basado en la Transformada de Fourier, en este caso la hipótesis planteada es verdadera, las wavelets packet Perceptual con Daubechies 4 constituyen una alternativa frente al método basado en la Transformada de Fourier, la resolución de espectro es mejor, las wavelets packet Perceptual con Daubechies 4 tienen igual complejidad computacional que la Transformada de Fourier pero el tiempo de ejecución es menor.

Los resultados muestran evidentemente que la hipótesis planteada al inicio de la presente investigación es verdadera.

A continuación se analizan los resultados de los métodos propuestos en wavelets con uno de los métodos más robustos existentes y que es utilizado ampliamente por los reconocedores comerciales.

- En las tablas 8.9, 8.10, 8.11 y 8.12 se compara el método de extracción de características MFCC con los métodos basados en las wavelets de Haar, Daubechies 4, Daubechies 6 y Coiflets 6 al aplicar la prueba ji-cuadrado de Mc Nemar para

datos correlacionados se se tiene un valor de $p < 0.01$ esto indica que existe una diferencia altamente significativa, En este caso el método MFCC es evidentemente superior , a pesar de que las wavelets tienen complejidad computacional mucho menor ($O(n)$ frente $O(n \log n)$), estos métodos no son muy atractivos por su basa tasa de reconocimientos

- En las tablas 8.13, 8.14, 8.15 y 8.16 se compara el método de extracción de características MFCC con los método basado en las wavelets Packet Walsh, wavelets Packet Daubechies 4, wavelets Packet Daubechies 6 y wavelets Packet Perceptuales con Daubechies 4 al aplicar la prueba ji-cuadrado de Mc Nemar para datos correlacionados se se tiene un valor de $p < 0.01$ esto indica que existe una diferencia altamente significativa, el método basado en wavelets que mejor se comporta es el basado en las wavelets packet Daubechies 6 seguido de las wavelet packet Daubechies 4, esto indica que los métodos propuestos podrían en un futuro mejorarse para así de esta manera poder alcanzar el grado de reconocimiento que tienen los MFCC. La mejora de las wavelets Packet esta en el menor numero de procedimientos empleados para obtener los valores de características, esto hace que tengan menor costo computacional que los MFCC.

Las tasas de reconocimiento y las tasas de error por método son descritas en las tablas 8.1, 8.2, 8.3 y 8.4; puede verse que existe un mayor grado de reconocimiento con el uso de la distancia de Chebyshev y una condición de Slope Constraint $p = 1$

siendo la técnica de extracción de características usando wavelets packets Daubechies 6 las que tienen mejor desempeño en comparación de las técnicas allí presentes, se muestra además su comparación frente a la técnica MFCC.

Es de evidenciarse la baja tasa de reconocimientos de las wavelets de Haar, Daubechies 4, Daubechies 6 y Coiflets 6 a pesar de su menor costo computacional.

En la tabla 8.5 se puede notar que las palabras : *salir, uno, cerrar y error* son las que tienen mayor índice de reconocimiento.

En algunas palabras como *tres, eliminar y terminar* tuvieron bajo reconocimiento pues se confundían con los valores de entrenamiento de *test y caminar*.

Parte IV

Conclusiones

Conclusiones

Al finalizar la presente investigación se concluyó lo siguiente:

1. El mejoramiento del espectro se da gracias al análisis tiempo frecuencia de las wavelets, con éstas podemos saber aproximadamente el aporte de las frecuencias por nivel de tiempo en las señales de habla, pues analiza con pequeñas wavelets componentes de alta frecuencia en la señal y con wavelets mas grandes componentes de baja frecuencia presentes en la señal, esto se traduce en la tasa de reconocimientos que proporcionan los wavelets
2. Una extracción de características usando solamente la Transformada de Fourier no dá buenos resultados, pues ésta hace un análisis tiempo frecuencia de la señal con ventanas del mismo tamaño para todos los niveles de frecuencia sacrificando resolución de tiempo o frecuencia según se enpequeñezca o agrande la ventana de análisis empobreciendo de esta manera la resolución espectral o temporal a diferencia de las wavelets que utilizan tamaños de ventanas diferentes según la frecuencia a analizar.
3. Existen diversos métodos utilizados para la extracción de características en el procesamiento digital de la señal para el Reconocimiento automático del Habla, siendo uno de los más robustos y usados el MFCC, el cual presenta diferentes formas de implementación. La efectividad de reconocimiento de habla

del MFCC radica en que es un buen modelo de representación de la producción y percepción de habla, el cual es obtenido gracias a la agrupación de diversos métodos como el cepstrum, la escala Mel, Transformada de Fourier, etc. El mismo hecho de agrupar varios métodos para obtener mayor efectividad en el reconocimiento, eleva el tiempo de ejecución del algoritmo, llegando a obtener una complejidad tiempo de $O(n \log n)$, por frame de tamaño n , y un tiempo de ejecución mucho mayor.

4. Los wavelets pueden ser utilizados alternativamente, para el procesamiento digital de la señal de habla. aprovechando el análisis que permiten y su rápida implementación computacional
5. Las wavelets discretas implementadas con el algoritmo de banco de filtros, para la extracción de características brindan una tasa de reconocimiento bajo, por lo cual se recurren a las wavelets packets que tienen mayor resolución en frecuencias, a las cuales hemos adaptado de tal manera de que en los espacios de resolución las wavelets que se comportan como bases, tengan una frecuencia aproximada a la de la escala Mel.
6. La complejidad computacional de los algoritmos de extracción de características usando las wavelets y las wavelets packets es de $O(n)$ y de $O(n \log n)$ respectivamente.

7. La complejidad computacional de los MFCC y de las wavelets packet es la misma $O(n \log n)$, pero el menor tiempo de ejecución corresponde a las wavelets packet, debido al menor número de procedimientos utilizados.
8. La ventaja de utilizar wavelets radica, en la variedad de funciones wavelet que se puede escoger, además de sus formas discretas y continuas.
9. Debido a la mejor resolución tiempo-frecuencia que tienen las wavelets, pueden analizar las señales de habla con menor número de procedimientos.
10. Las wavelets que mejor funcionan, son aquellos que tienen su espectro parecido a un filtro paso de banda ideal.
11. Tomar las energías de los bandos de frecuencia y luego sacar el cepstrum de la energías no dan resultados buenos frente a las wavelets esto se debe a que el análisis mediante la Transformada de Fourier es semejante a un vector, mientras que el análisis mediante la transformada wavelet es más parecida a un árbol donde los valores presentes en cada hoja corresponde a los coeficientes en cierto nivel de resolución y cierto nivel de frecuencia, esto es lo que permite decir que las wavelets permiten un mejoramiento del espectro en la señal de habla
12. Finalmente se concluye que la hipótesis planteada es cierta, el análisis mediante wavelets sí constituye una alternativa frente a un análisis de Fourier, mejora la resolución de espectro y el tiempo de ejecución en la extracción de características,

y su tasa de reconocimientos si bien es cierto no supera a la técnica de los MFCC, muestran resultados interesantes que podrían ser útiles para posteriores investigaciones.

Para trabajos futuros podemos tener las siguientes consideraciones:

Se pueden utilizar otra gamma de wavelets teniendo en consideración los resultados obtenidos en la presente tesis, como también la utilización de otros wavelets continuos.

Emplear otra manera de obtener las características como podría ser una técnica basada en la ventaja del análisis tiempo frecuencia de las wavelets , esto es segmentar la señal de una manera no uniforme , con tamaños variables de frames obtenidos gracias a las variaciones locales en la señal.

La tasa de reconocimiento puede variar según el reconocedor que se use, en este caso hemos hecho uso de un reconocedor basado en un algoritmo optimizado de programación dinámica DTW, pero puede utilizarse reconocedores más avanzados como es los basados en Redes Neuronales , en el uso de probabilidades como son los Modelos Ocultos de Markov, Redes Bayesianas, etc

Utilizar modelos en lugar de plantillas, para así disminuir aún mas el tiempo de ejecución, y la tasa de error.

La tasa de reconocimiento que se ha mostrado en los métodos implementados se ha obtenido construyendo un reconocedor del tipo independiente del hablante, tomando como plantillas las muestras de cinco personas, y para las pruebas con 60 personas;

esta tasa puede variar muy notablemente en un sistema dependiente del hablante, donde es de esperarse tasas de reconocimientos mucho mayores.

Comentarios

Como comentario final se debe agregar el análisis hechos por las wavelets continuas; éstas analizan las señales de tal manera que los parámetros de traslación y escalamiento a, b toman cualquier valor real ($a \neq 0$); hay muchas funciones wavelets contínuas en este caso decidimos probar con la wavelet contínua de Morlet, por tener esta un espectro parecido a un filtro pasa banda, los detalles de su implementación y algunas comparaciones se detallan a continuación.

La Wavelet de Morlet llamada también Wavelet Gaussiano Modulado o Wavelet de Gabor, esta función propuesta por Gabor y ampliamente usada por dos científicos : Morlet y Kronland. está definida como la Trasformada de Fourier de una gaussiana trasladada de tal manera que en promedio la función sea cero.

$$\psi(x) = \pi^{-\frac{1}{4}}(e^{-i\omega x} - e^{-\frac{\omega^2}{2}})e^{-\frac{x^2}{2}} \quad (9.1)$$

y una respuesta en el dominio de la frecuencia.

$$\psi(\omega) = \pi^{-\frac{1}{4}}[e^{-(\omega-\omega_0)} - e^{-\frac{\omega^2}{2}}e^{-\frac{\omega_0^2}{2}}] \quad (9.2)$$

la extracción de características fué hecha como sigue: primero se construyéron 32 wavelets de Morlet en diferentes escalas, para lograr esto se implementó la siguiente wavelet de Morlet

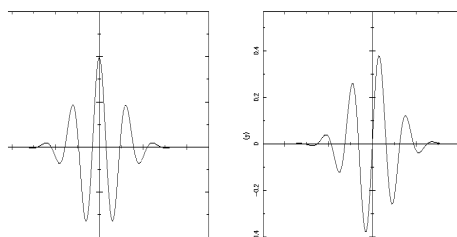


Figura 9.1 Parte real e imaginaria de las Wavelets de Morlet

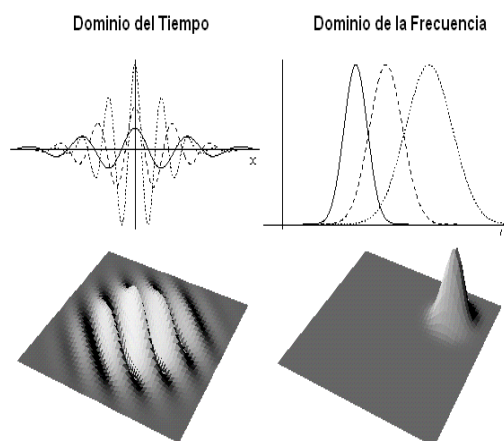


Figura 9.2 Parte real de las Wavelets de Morlet en el dominio del tiempo y en el dominio de la frecuencia en 1D y 2D respectivamente

$$\psi(x) = \pi^{-\frac{1}{4}} (e^{-i\omega x} - e^{-\frac{\pi^2 \alpha^2}{4}}) e^{-\frac{x^2}{\alpha^2}} \quad (9.3)$$

donde α controla el tamaño de la envoltura gaussiana.

Luego en la implementación para evitar una convolución muy costosa en el dominio de las wavelets con la señal de habla se procedió a calcular la Transformada Inversa de Fourier de la multiplicación en el dominio de la frecuencia de las wavelet de Morlet y del segmento de la señal de habla, esto hace que el procedimiento a este punto tenga ya una complejidad computacional de $O(n \log n)$; en consecuencia la implementación

de este tipo de wavelets no mejora la complejidad computacional, pero la tasa de reconocimientos es comparable a la técnica de los MFCC.

Para obtener las características se procedió a hacer algo similar a los bins de los MFCC , pero en lugar de filtros triangulares se usó una ventana gaussiana en dos dimensiones ,teniendo en cuenta a la escala Mel, y posteriormente se sacan los logaritmos de las energías aplicando finalmente una transformada discreta del coseno, estos serán los valores de características.

A continuación algunos resultados en donde se puede ver la alta tasa de reconocimiento de las wavelets de Morlet.

<i>Método</i>	<i>Tasa aceptación</i>	<i>Error</i>
Coefficientes Cepstrales en Escala Mel	85.32%	14.68%
Wavelet Continuo de Morlet	87.32%	12.68%

Tabla 9.1 Datos obtenidos utilizando la técnica de DTW como reconocedor, con distancia Chebyshev y con Slope Constrain P=1. Se observa la mejor performance en las Wavelet Continuos de Morlet

La aplicación de wavelets continuos, como el caso de las wavelets de Morlet , ayuda a incrementar la tasa de reconocimientos, esto es por que en el dominio de la frecuencia las wavelets son en realidad filtros de pasa banda, por el hecho de ser continuos podemos emplearlos como filtros para analizar solo frecuencias deseadas, esto es muy similar a los filtros triangulares (bins) utilizados en la técnica del MFCC, es por eso que con wavelets continuos es de esperarse tasas de reconocimientos similares

que a las tasas de los MFCC, estos wavelets continuos para evitar una complejidad computacional mucho mayor en la implementación en el proceso de convolución con la señal de habla, es mejor llevarlos al dominio de la frecuencia, y multiplicarlos con la señal en el dominio de la frecuencia, luego aplicar una transformada inversa de Fourier, esto nos da una complejidad de $O(n \log n)$.

Referencias Bibliográficas

- [Aboufadel, 2001] Aboufadel, E. (2001). A wavelets approach to voice recognition. *Grand Valley State University*.
- [Atal and Schroeder, 1968] Atal and Schroeder (1968). Predictive coding of speech signals. *Report of the 6th Int. Congress on Acoustics, Tokio, Japan*.
- [Barrantes and Rodrigo, 2001] Barrantes E., Rodrigo (2001). Investigación: Un camino al conocimiento, un enfoque cualitativo y cuantitativo. *San Jose, CR.: EUNED..*
- [Baun and Eagon, 1967] Baun, L.E. and J.A. Eagon (1968). "AAAn Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology," *Bulletin of American Mathematical Society, 1967, 73, pp. 360-363*.
- [Beng, 2000] Beng (2000). The use of wavelet transforms in phoneme recognition. *Dept. of Electrical and Computer Engineering, The University of Newcastle Australia*.
- [Bernal, 2000] Bernal, J., J. Bobadilla and P. Gómez (2000). Reconocimiento de Voz y Fonética Acústica. *Dept. de informatica, Universidad Politécnica de Madrid*.
- [Chu and George, 2000] Chu, E.,A. George (2000). Inside the FFT Black Box Serial and Parallel Fast Fourier Transform Algorithms. *CRC Press Boca Raton London New York Washington, D.C*.
- [Coifman Meyer and Wickerhauser, 1992] Coifman, R.R.,Y. Meyer and M.V. Wickerhauser (1992). Wavelet analysis and signal processing. *In Wavelets and their Applications, pages 153-178, Boston, 1992. Jones and Barlett. E. Ruskai et al. editors..*
- [Cooley and Tukey, 1965] Cooley, J.W.,J.W. Tukey (1965). An algorithm for the machine calculation of complex Fourier Series. *Math. Comp. 19:297-301,1965*.
- [Daubechies, 1992] Daubechies, I. (1992). Ten lectures on wavelets. *Society for Industrial and Applied Mathematics*.
- [Davis and Mermelstein, 1980] Davis, S. and P. Mermelstein (1980). Comparison of Parametric Representation for Monosyllable Word Recognition in Continuously Spoken Sentences. *IEEE Trans on Acustics, Speech and Signal Processing, 1980, 28(4),pp. 357-366..*

- [Gentleman and Sande, 1966] Gentleman, W.M.,G.Sande (1966). Fast Fourier Trasforms - for fun and profit. *Fall Joint Computer Conf., AFIPS Proc., Vol. 29, pp. 563578. Washington, D.C., Spartan, 1966..*
- [hermansky, 2005] hermansky, H. (2005). Perceptual linear predictive analysis of speech. *J. Acoust. Soc. Am.*
- [Huang and Hon, 2001] Huan, X., A. and Hon, H. (2001). Spoken language processing a guide to theory, algorithm and system development. *Prentice Hall, New Jersey.*
- [Jelinek, 1998] Jelinek (1998). Statistical methods for speech recognition, language, speech and communication. *Cambridge, MA, MIT Press.*
- [M. Siafarikas, 2000] M. Siafarikas, Todor Ganchev, N. F. (2000). Objctive wavelet packet features for speaker verication. *Wire Communications Laboratory, University of Patras, Greece.*
- [Mallat, 1989] Mallat. (1989). Multiresolution approximation and wavelets. *Trans. Amer. Math. Soc., 315, pp. 69-68..*
- [Mantha, 1998] Mantha, V., R. Y. a. J. (1998). Implementation and analysis of speech recognition frontends. *Departament of Electrical and Computer Engineering, Mississippi State University.*
- [Markowitz, 1996] Markowitz, J. (1996). Using Speech Recognition. *Upper Saddle River, Prentice Hall..*
- [Meyer, 1986] Meyer. (1986). Ondelettes, fonctions splines et analyses graduées. *Lectures given at the University of Torino, Italy..*
- [Pérez, 2000] Pérez, C. (2000). Técnicas de Muestreo Estadístico *ED. grupo Editor Alfaomega, México..*
- [Ruhi Sarikaya and Hansen, 2001] Ruhi Sarikaya, B. L. P. and Hansen, J. H. L. (2001). Wavelet packet transform features with aplication to speaker identification. *Robust Speech processing Laboratory, Duke University.*
- [Sakoe and Chiba, 1978] Sakoe, H. and Chiba, S. (1978). Dynamic programmig algorithm optimization for spoken word recognition. *IEEE.*
- [Sarikaya and Hansen, 2000] Sarikaya, R. and Hansen, J. (2000). High resolution speech features parametrization for monophone based stressed speech recognition packet transform features with aplication to speaker identification. *IEEE Signal Processing Letters, Vol.7,NO.7,July 2000.*

Apéndice

Prueba JI Cuadrado de MC Nemar-Datos Correlacionados

Uno de los usos de esta prueba es que permite contrastar los resultados de una prueba o método propuesto con los que se obtengan con una prueba diferente denominada *Prueba Estándar* o *Prueba de Oro*, a la cual se asigna un valor de certeza diagnóstica. Los resultados pueden presentarse de la siguiente manera:

Procedimiento de Prueba

a *Formulación de Hipótesis*

$$H_0 : P_1 = P_2$$

$$H_1 : P_1 \neq P_2$$

(0.4)

P_1 : Proporción de aciertos con el método 1

P_2 : Proporción de aciertos con el método 2

$$P_1 = \frac{a+b}{n}$$

$$P_2 = \frac{a+c}{n}$$

b Nivel de Significación:

$$\alpha = 0.05$$

c Estadística de Prueba

$$Z = \frac{b-c}{\sqrt{b+c}}$$

$$Z^2 = \chi_{MN} \text{ con un grado de libertad}$$

$$\chi^2 = \frac{(b-c)^2}{b+c} \rightarrow \chi_1^2$$

d Valor Tabular

$$\chi_{tab}^2 = \chi_{1;1-\frac{\alpha}{2}}^2$$

e Criterio de Decisión

- Si $p < 0.05$, Existe diferencia estadísticamente significativa entre ambos métodos.
- Si $p < 0.01$, Existe diferencia altamente significativa entre ambos métodos.
- Si $p > 0.05$, No existe diferencia estadísticamente significativa entre ambos métodos

LORITO 3.14

Se implementó un software para pruebas al que se le denominó LORITO v 3.14

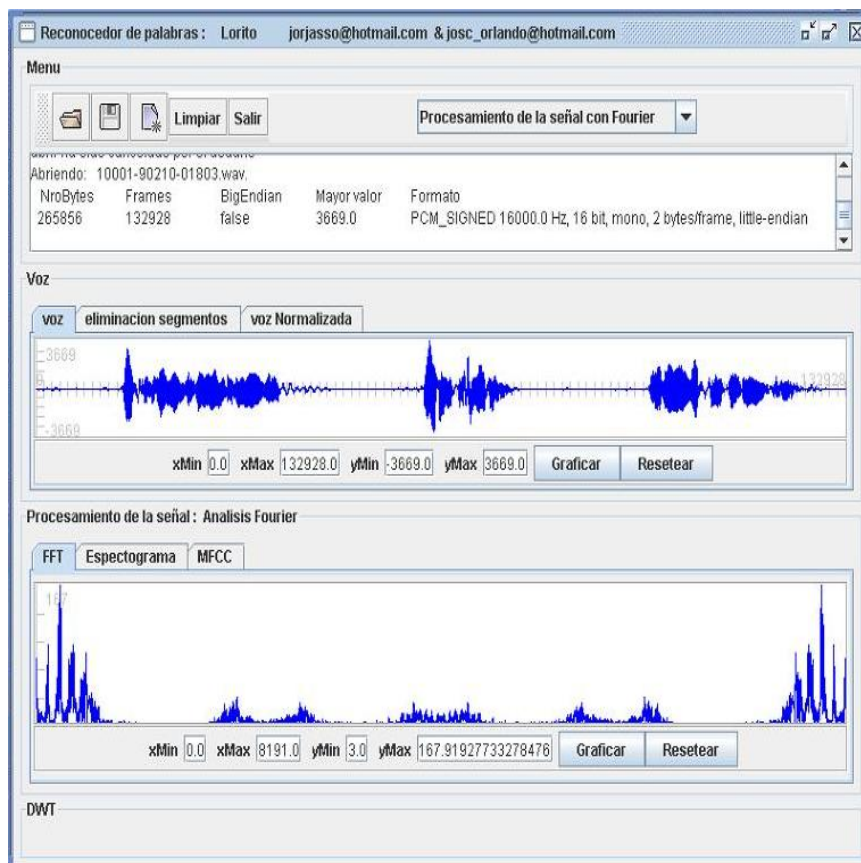


Figura 0.3 Lorito graficando la Transformada Discreta de Fourier de una señal de habla.

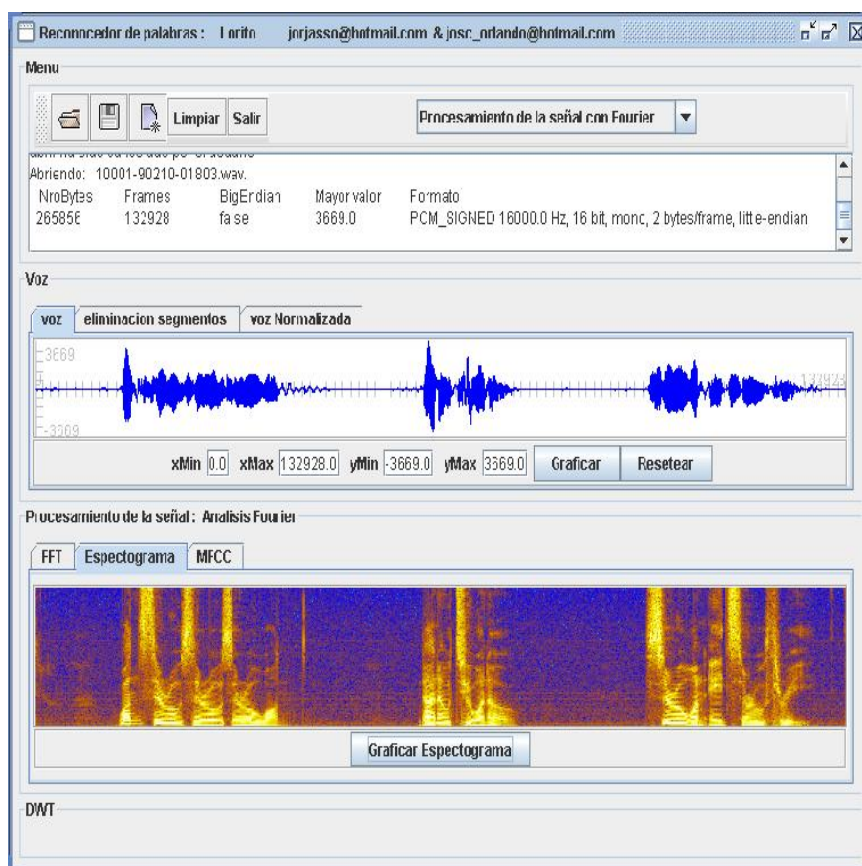


Figura 0.4 Lorito graficando el espectrograma como paso previo para la extracción de características MFCC.

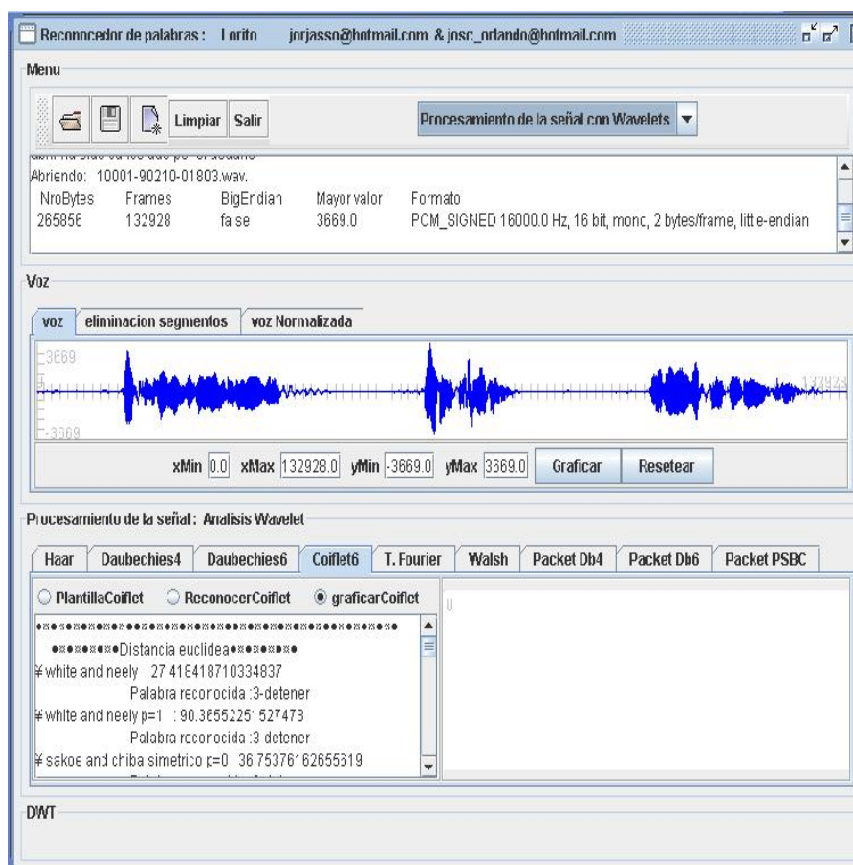


Figura 0.5 Lorito mostrando resultados de reconocimiento por medio de las wavelets.

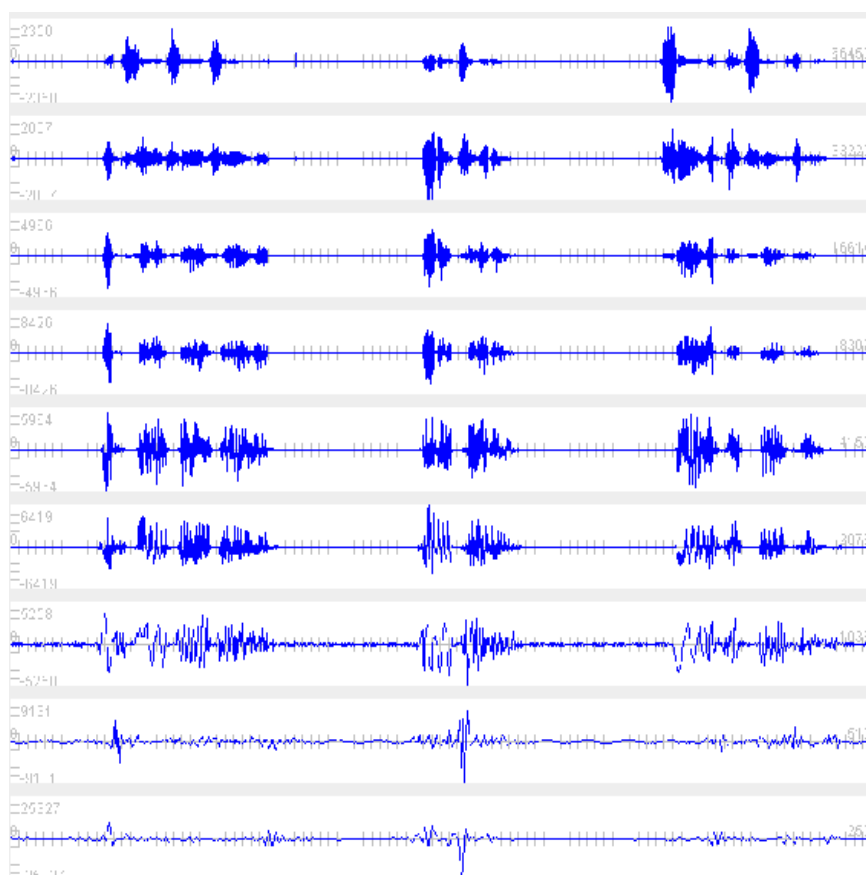


Figura 0.6 Lorito graficando los coeficientes en diferentes espacios de resolución de una señal de habla .