UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

ESTUDIO DE TÉCNICAS DE SELECCIÓN DE BANCOS DE FILTROS INDUCIDOS POR WAVELET PACKETS PARA EXTRACCIÓN DE CARACTERÍSTICAS EN RECONOCIMIENTO DE VOZ

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRICISTA

EDUARDO HERNÁN PAVEZ CARVELLI

PROFESOR GUÍA:

SR. JORGE SILVA SÁNCHEZ

MIEMBROS DE LA COMISIÓN:

SR. HÉCTOR AGUSTO ALEGRÍA

SR. NÉSTOR BECERRA YOMA

SANTIAGO DE CHILE
AGOSTO 2011

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELECTRICISTA
POR: EDUARDO HERNÁN PAVEZ CARVELLI
FECHA: AGOSTO DE 2011
PROF. GUÍA: SR. JORGE SILVA SÁNCHEZ

ESTUDIO DE TÉCNICAS DE SELECCIÓN DE BANCOS DE FILTROS INDUCIDOS POR WAVELET PACKETS PARA EXTRACCIÓN DE CARACTERÍSTICAS EN RECONOCIMIENTO DE VOZ

En procesamiento de señales y reconocimiento de voz, uno de los principales tópicos es la extracción de características. Dada una señal digitalizada esta se transforma a una más compacta de acuerdo a algún criterio dependiente del problema. Por ejemplo, la voz es extremadamente redundante, y contiene información de múltiples fuentes como identidad del locutor, estado emocional y secuencia fonética. En reconocimiento de voz, se desea extraer características que preserven la discriminabilidad entre unidades acústicas pero tengan menor dimensión que la señal original.

En esta memoria se estudiarán metodologías de extracción de características para reconocimiento de voz utilizando criterios de discriminabilidad entre unidades fonéticas. Como punto de partida se considerará la técnica de extracción de características MFCC utilizada en los reconocedores estándar y con ellas se entrenará un reconocedor con el software HTK para la base de datos TIMIT.

Los MFCC se implementan con bancos de filtros; usando esa motivación y aprovechando la gran colección de formas de representar señales que permite la transformada Wavelet Packet (WP), se propuso un método de selección de bancos de filtros WP considerando discriminación entre unidades acústicas. La memoria se centrará en aplicar esta metodología y las soluciones que esta entrega para proponer una alternativa a los MFCCs. La forma de evaluar los rendimientos de los distintos métodos será mediante el porcentaje de reconocimiento fonético en un subconjunto de test de TIMIT.

La parte teórica del trabajo consiste en estudiar propiedades y formas de aplicar los WP, y como se construyen los sistemas de reconocimiento fonético. En la práctica, en implementar la transformada WP y los WPCC en C++, implementar reconocedores fonéticos en el soft-

ware HTK, y unir todos estos bloques usando el lenguaje PERL, para correr experimentos en distintos escenarios.

El aporte de este trabajo son las características Wavelet Packet Cepstral Coefficients (WPCC), se estipulan criterios concretos de diseño para los bancos de filtros WP, con el objetivo de obtener la mayor información para discriminar fonemas. Los WP obtenidos tienen alta selectividad en frecuencia y una estructura que privilegia la resolución en bajas frecuencias principalmente el rango [200Hz-1000Hz] del espectro acústico.

Índice general

Capítulo 1

Introducción

1.1. Antecedentes

La extracción de características es una de las etapas mas importantes en el diseño de sistemas automáticos de reconocimiento de voz [1]. La metodología mas utilizada para la extracción de características acústicas son los *Mel-frequency Cepstral coefficients* (MFCC). Los MFCCs corresponden a un esquema de análisis segmental en el tiempo, en el cual coeficientes de energía del espectro son extraídos de un banco de filtros, con frecuencias centrales proyectadas uniformemente en la escala Mel [1]. Esta escala se deriva de estudios del sistema auditivo humano.

Como alternativas, se han propuesto nuevas técnicas de procesamiento de señales para realizar este análisis segmental [2, 3, 4, 5, 6], donde el uso de Wavelets y Wavelet Packets [7, 8, 9] ha sido de particular interes en este contexto.

Los Wavelet Packets (WP) [9, 8, 10] han sido ampliamente aplicados a esquemas de compresión, detección y clasificación [11, 12, 13, 14, 15, 16, 17]. Estas herramientas son particularmente apropiadas para el análisis de procesos de series de tiempo pseudo-estacionarios, tales como el proceso de producción acústico de la voz [2, 4, 18, 16].

WPs son un tipo de bases estructuradas, donde los vectores ortogonales que componen la base son generados mediante una cantidad finita de transformaciones elementales [9, 7, 13]. Donde desde un punto de vista ingenieril, estas representaciones son atractivas porque pueden implementarse mediante un bloque básico de filtros y operadores de sub-muestreo [9]. Pueden ser usados para caracterizar un amplio tipo de descomposiciones del espacio de señales, y en particular, proveen un modo de generar particiones tiempo-frecuencia del espacio de obser-

vaciones.

En conclusión, los WP inducen una familia de bancos de filtros estructurados, con una variedad de representaciones tiempo-frecuencia, con el potencial de enriquecer el modo convencional en que los MFCC describen el comportamiento segmental del proceso acústico de producción de la voz.

Los WP y bancos de filtros multi-tasa, han sido utilizados como alternativas a los MFCC en reconocimiento de voz [3, 4, 5, 6]. En particular Farooq *et al.* [3] propuso una representación mediante WP, en la cual el banco de filtros usado para su implementación imitaba la partición en frecuencia de la escala Mel. Con los filtros de *Daubechies* (DB)[7], muestran mejoras en la clasificación de subcategorías fonéticas en una parte del corpus TIMIT. Mas recientemente, Choueiter *et al.* [4] exploró el problema de diseño de banco de filtros de dos canales, y en particular el nuevo esquema de bancos de filtros racionales. El foco de este trabajo era mejorar la selectividad en frecuencia con respecto a los filtros de Daubechies, y obtener mejores resultados al considerar una partición con estructura similar a la escala Mel y no la típica particion de Wavelets tradicionales. En este escenario se obtuvieron mejores resultados con respecto a los MFCCs en una tarea simplificada de clasificación de segmentos fonéticos.

Estos trabajos presentan evidencia concreta de las ventajas de utilizar Wavelets y Wavelet Packets para parametrizar la señal acústica de voz. Sin embargo, el problema de adaptar la base WP a un problema particular, en el sentido de encontrar la topología de banco de filtros, dentro de la colección de bases WP, que capture mejor la información tiempo-frecuencia para este problema de reconocimiento de patrones no ha sido explorado.

Además, los resultados reportados hasta ahora solo han considerado escenarios experimentales simplificados, en términos del problema de clasificación o de la base de datos. Por esto un análisis sistemático, con experimentos de reconocimiento fonéticos estandarizados, podría ser beneficial para validar el uso de características WP, como una alternativa competitiva a los MFCC.

1.2. Descripción del Problema

En este trabajo se propone el uso de Wavelet-Packet Cepstral Coefficient (WPCC), y se muestran resultados experimentales que complementan la validez del uso de WP en Extracción de Características para Reconocimiento de Voz. Se construye la metodología sobre

las ideas recientemente propuestas por Silva *et al.* [2], en las cuales el problema de selección optima de bancos de filtros para reconocimiento de patrones fue formulado basándose en el principio de la mínima probabilidad de error de decisión.
Se exploran metodologías de selección de bancos de filtros WP para proponer una familia de WPCC. Estas características son basadas en log-energías, procesadas con la Transformada Coseno Discreta *Discrete Cosine Transform, (DCT)*, (Cepstrum), según lo propuesto en [3], donde las energías son obtenidas de un banco de filtros seleccionado de la familia de WP.
Para la selección de banco de filtros, se utiliza un criterio de regularización adoptado de la literatura de selección de bases estructuradas como árbol [2, 12, 19, 20]. En particular, se usan métodos basados en *Energia* acústica, el *Fisher-scatter ratio*[21], y la Divergencia de Kullback-Leibler (KLD), como medidas de fidelidad. Los últimos dos criterios consideran la discriminación fonética, mientras que el basado en energía, aumenta la resolución en frecuencia en las bandas con mayor energía acústica, propuesto en [18] para el problema de clasificación de texturas.
Como resultados complementarios, se corre un experimento de reconocimiento fonético estándar en el corpus TIMIT.En el que se contrastan las diferentes soluciones con respecto a varios elementos de diseño. Entre ellos están las medidas de fidelidad, el número de bandas, número de características y la selectividad en frecuencia del filtro de dos canales que induce la familia de WP.

1.3. Objetivos

Los objetivos planteados como fundamentales en esta memoria son los siguientes:

- Diseñar una metodología de Extracción de Características para Reconocimiento de Voz, basada en Wavelet Packets, los Wavelet Packet Cepstral Coefficients (WPCC).
- Aprovechar la estructura de los WP y las técnicas conocidas para su selección óptima, y así escoger la mejor representación de acuerdo a criterios de discriminación fonética.
- Evaluar las características WPCC en un problema de reconocimiento de series de tiempo fonética utilizando un reconocedor de voz estándar basado en *Hidden Markov Models* (HMMs).

- Evaluar los WPCC en función de variables como selectividad en frecuencia y criterio de selección de WP.

1.4. Estructura de la Memoria

La memoria se estructura de la siguiente forma. El capítulo 2 trata la extracción de características, reconocimiento de voz e introduce los WPCC. El capítulo 3 presenta el material de referencia para entender las propiedades de bancos de filtros de los WP y el 4 como seleccionarlos para este problema de clasificación. Finalmente los capítulos 5 y 6 muestran las estructuras de banco de filtros obtenidas, y resultados de reconocimiento fonético, respectivamente. Las conclusiones se muestran en el capítulo 7, y finalmente un Anexo.

Capítulo 2

Parametrización y Reconocimiento de Voz

La información contenida en la voz puede ser un mensaje, identidad de la persona, estado emocional, etc. de naturaleza discreta, pero se transmiten mediante una señal de naturaleza continua.
En reconocimiento de voz el interés está en el mensaje (secuencia fonética) y por lo tanto se debe representar la señal en la forma mas compacta que contenga esa información. La representación de la señal en reconocimiento de patrones se denomina Extracción de Características, la cual busca reducir la dimensionalidad de la señal, preservando la información relevante. En los últimos años, MFCCs han sido adoptados como la representación acústica estándar y están basados en estudios de percepción humana.
La modelación se hace por unidad acústica básica , y las utilizadas en reconocimento de voz son los fonemas. Es sabido que mejoras en el reconocimiento fonético se traducen en mejoras en reconocedores de voz basados en esos modelos. La problemática del reconocimiento de voz, se puede separar en tres tareas fundamentales:

- Extracción de Características
- Entrenamiento de modelos
- Reconocimiento

2.1. Sistema de Reconocimiento de Voz

Los sistemas actuales de reconocimiento de voz, se basan en la teoría de reconocimiento estadístico de patrones. La señal de voz es convertida en una secuencia de vectores acústicos X (extracción de características) y por otra parte la señal acústica es consistente con un mensaje, o secuencia de palabras W. Entonces el trabajo de un sistema de reconocimiento de voz, es determinar la secuencia de palabras más probable $\hat{W}$, dados los vectores acústicos u observaciones. Para lograrlo, se utiliza la regla de Bayes para descomponer la probabilidad $P(W|X)$ de la forma

$$\hat{W} = \arg\max_{W} P(W|X) \tag{2.1}$$

$$= \arg\max_{W} \frac{P(W)P(X|W)}{P(X)} \tag{2.2}$$

El termino del denominador es independiente de la secuencia de palabras, por lo que el problema de optimizacion solo depende de los términos $P(W)$ y $P(X|W)$. El primero corresponde a la probabilidad a priori de observar una secuencia W, independiente de la señal acústica observada, y esta probabilidad se determina mediante un modelo de lenguaje, que en su forma mas simple, considera las dependencias solo de las palabras vecinas (modelo de Markov de primer orden) y se denomina modelo de bi-grama.

El segundo término representa la probabilidad de observar un vector X, dada una secuencia de palabras específica. Esta probabilidad está dada por los modelos acústicos. Los modelos acústicos mas utilizados son los denominados Hidden Markov Models (HMM), que representan unidades básicas tales como fonemas contexto-independientes (mono-fonemas) o fonemas contexto-dependientes (bi-fonemas y tri-fonemas), luego la concatenación de HMMs da lugar a palabras y secuencias de palabras.

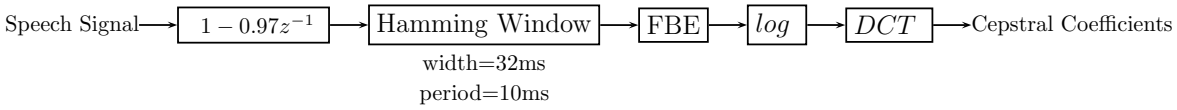
Para aplicar esta metodología a sistemas prácticos, se requiere solucionar varios problemas. La extracción de características debe ser capaz de extraer de la señal de voz toda la información acústica necesaria, de la forma mas compacta y compatible con los modelos basados en HMMs. Además los HMMs deben representar de forma precisa las distribuciones de los sonidos en todos los contextos, considerando además que estas deben ser estimadas de los datos. Finalmente el modelo de lenguaje debe ser capaz de predecir correctamente las palabras en función de sus predecesoras.

La extracción de características es abordada en profundidad en esta memoria, mientras que la implementación del reconocedor basado en HMMs se hace con el software HTK[22].

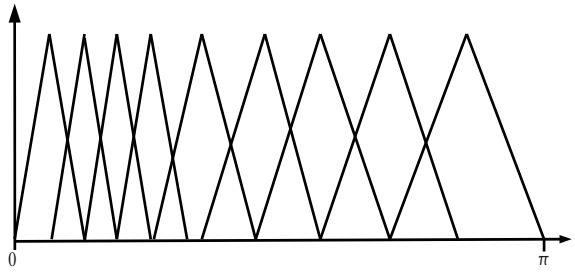
2.2. Extracción de Características

La información contenida en la voz como identidad del locutor o mensaje pueden ser analizadas desde el punto de vista del ser humano o del sistema de reconocimiento automático. Algunas características identificables por las personas son "claridad", "aspereza" y "volumen". Otras de más alto nivel son prosodia, tono, velocidad de articulación y dialecto. Estas características son más fácilmente identificables por el ser humano que por las máquinas. Por otra parte, las características de naturaleza acústica son mejor medidas por sistemas computacionales, tales como espectro de la voz o frecuencias fundamentales.

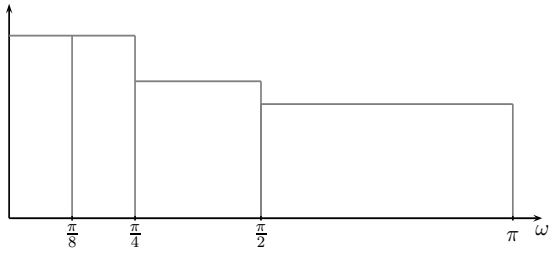
El proceso de extracción de características para reconocimiento de voz, y mas generalmente el proceso de codificación se realiza de forma segmental. La técnica estandar para realizar extracción de características se basa en energías obtenidas de un banco de filtros procesadas por la transformada cepstral [1], (Figura 2.1a). Dada la señal acústica, el esquema sigue las siguientes fases: un filtro pasa alto $1 - 0{,}97z^{-1}$ de pre-énfasis en la señal completa que compensa la radiación de los labios; segmentación con una ventana de Hamming de tamaño $32ms$, creando segmentos acústicos traslapados cada $10ms$ para el; cada segmento es pasado a través de un banco de filtros triangulares con frecuencias centrales en la escala Mel (figura 2.1b); finalmente, en cada vector de energías se aplica la función logaritmo y la transformada coseno discreta (DCT), para crear los Mel Frequency Cepstral Coefficients (MFCC)[23].



(a) Algoritmo de Extracción de Características



(b) Banco de filtros Mel



(c) Banco de filtros tipo wavelet

Figura 2.1: Paradigma estándar de parametrización de la señal acústica para reconocimiento de voz.

En este trabajo, se explora una extensión a esta metodología de extracción de características, donde en vez de usar el banco de filtros de la escala Mel, se estudia la extensa colección de bancos de filtros inducidos por las bases Wavelet Packet [9, 24]. Y asi reemplazar este banco de filtros de forma natural en el esquema de la figura 2.1a y dar lugar a los Wavelet Packet Cepstral Coefficients (WPCC).
Las siguientes secciones explican la metodología empleada para derivar estructuras de WP mejor adaptadas a la discriminación de fonemas, partiendo con la selección de banco de filtros y la aplicación de los WPCCs a un problema de reconocimiento de voz (*Continuous Speech Recognition*).

Capítulo 3

Wavelet Packets

Para generalizar la Transformada Wavelet, Coifman, Meyer y Wickerhauser [25] propusieron Wavelet Packets (WPs), una familia de bases con una estructura de árbol binario, donde la transformada Wavelet corresponde a un caso particular de WP. Esta familia ofrece diferentes tipos de representación tiempo-frecuencia, y consecuentemente tiene el potencial de adaptarse a complejos tipos de series de tiempo, como el proceso acústico de producción de voz. A continuación se presenta una introducción a esta, con énfasis en su implementación mediante bancos de filtros. Detalles pueden encontrarse en [24, 9].

3.1. Descomposición binaria en sub-espacios

Los Wavelet Packets pueden entenderse como una colección de bases indexadas, que ofrecen una descomposición binaria del espacio de señales al dividir su contenido en frecuencia [24]. Sea $\mathbb{X}$ el espacio de señales (asociado a un *nivel finito de escala* 2^L o de *resolución* 2^{-L}), con una base ortonormal $\mathbf{B}_L \equiv \left\{\phi_L(t-2^L n)\right\}_{n\in\mathbb{Z}}$. El esquema de los WPs se basa en que es posible descomponer $\mathbf{B}_L$ en dos bases ortonormales $\mathbf{B}^0_{L+1} \equiv \left\{\phi^0_{L+1}(t-2^{L+1}n)\right\}_{n\in\mathbb{Z}}$ y $\mathbf{B}^1_{L+1} \equiv \left\{\phi^1_{L+1}(t-2^{L+1}n)\right\}_{n\in\mathbb{Z}}$, donde denotando como $U^p_{L+1} \equiv span\left\{\mathbf{B}^p_{L+1}\right\}$, para $p\in\{0,1\}$ se tiene $\mathbb{X} = U^0_{L+1}\bigoplus U^1_{L+1}$ [24].
La estructura de los WPs viene dado por la descomposición en sub-espacios $\mathbf{B}^1_{L+1}$ y $\mathbf{B}^0_{L+1}$, y en que estos son inducidos por un par de filtros discretos del tipo *conjugate mirror filters*(CMF)

(ec. 3.3)[24, Chap. 7.1.3], mas precisamente,

$$\phi^0_{L+1}(t) = \sum_{n=-\infty}^{\infty} h(n) \cdot \phi_L(t - 2^L n) \tag{3.1}$$

$$\phi^1_{L+1}(t) = \sum_{n=-\infty}^{\infty} g(n) \cdot \phi_L(t - 2^L n), \tag{3.2}$$

donde los CMF antes mencionados $h(n)$ and $g(n)$ están relacionados entre si por la propiedad de reconstrucción perfecta (ec. 3.4) [25], [24, Theorem 8.1],

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2 \tag{3.3}$$

$$g(n) = (-1)^{1-n} h(1 - n), \ \forall n \in \mathbb{Z} \tag{3.4}$$

En otras palabras, los CMFs mapean $\mathbf{B}_L$ en una base ortonormal alternativa $\mathbf{B}^0_{L+1} \cup \mathbf{B}^1_{L+1}$ para $\mathbb{X}$. Además podemos relacionar los contenidos en frecuencia de los subespacios U^0_{L+1} and U^1_{L+1} con el de $\mathbb{X}$ con la relación,

$$\hat{\phi}^0_{L+1}(\omega) = \hat{h}(2^L\omega) \cdot \hat{\phi}_L(\omega), \quad \hat{\phi}^1_{L+1}(\omega) = \hat{g}(2^L\omega) \cdot \hat{\phi}_L(\omega) \tag{3.5}$$

donde $\hat{\phi}^0_{L+1}(\omega)$ y $\hat{h}(2^L\omega)$ denotan la *Transformada de Fourier* (FT) y la *Transformada Discreta de Fourier*(DTFT) de $\phi^0_{L+1}(t)$ y $h(n)$ (igualmente, $\phi^1_{L+1}(t)$ y $g(n)$), respectivamente.
Las figuras 3.4 y 3.3 muestran la partición perfecta del contenido en frecuencia del espacio de señales que se obtiene con los *Wavelets de Shannon* (implementados con filtros ideales o de Shannon), para ilustrar la resolución en frecuencia del proceso binario de descomposición en sub-espacios mediante WPs. Iterando la aplicación del par de CMF $(h(n), g(n))$ en cada elemento de la base $\phi^0_{L+1}(t)$ y $\phi^1_{L+1}(t)$ (ver [24, Teorema 8.1]) se pueden construir (como árbol binario) bases alternativas y particiones en sub-espacios para $\mathbb{X}$.
Mas precisamente después de cierto número de descomposiciones, se puede obtener $\phi^p_j(t)$ para cualquier $j > L$ y para todo $p \in \{0, .., 2^{j-L} - 1\}$, donde $U^p_j = span\left\{\phi^p_j(t - 2^j n) : n \in \mathbb{Z}\right\}$, ver figura 3.1a. Entonces por construcción se tiene que , $U^p_j = U^{2p}_{j+1} \bigoplus U^{2p+1}_{j+1}$, $\forall j > L$, $\forall p \in \{0, .., 2^{j-L} - 1\}$, donde

$$\phi^{2p}_{j+1}(t) = \sum_{n=-\infty}^{\infty} h(n) \cdot \phi^p_j(t - 2^j n) \tag{3.6}$$

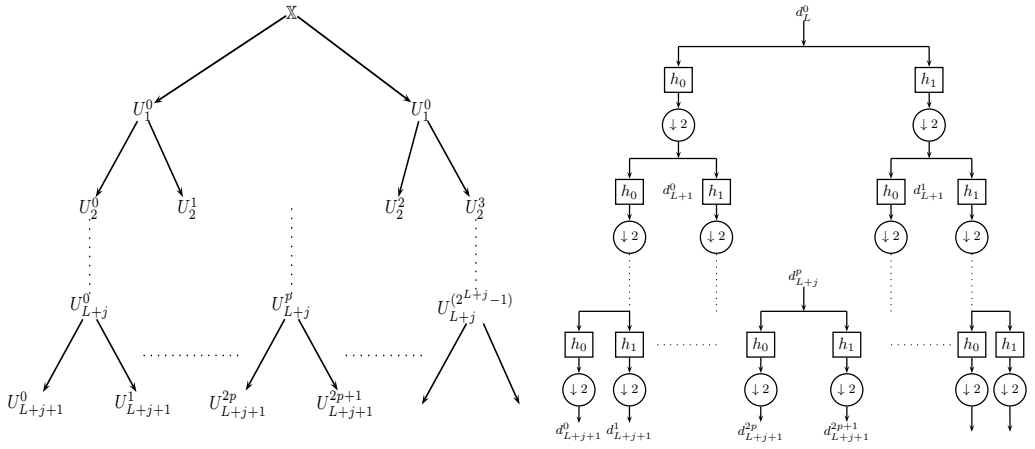
$$\phi^{2p+1}_{j+1}(t) = \sum_{n=-\infty}^{\infty} g(n) \cdot \phi^p_j(t - 2^j n) \tag{3.7}$$

La propiedad de descomposición en frecuencia también se deduce de la ecuación 3.5, es decir,

$$\hat{\phi}^{2p}_{L+j+1}(w) = \hat{h}(2^L w) \cdot \hat{\phi}^{p}_{L+j}(w) \tag{3.8}$$

$$\hat{\phi}^{2p+1}_{L+j+1}(w) = \hat{g}(2^L w) \cdot \hat{\phi}^{p}_{L+j}(w) \tag{3.9}$$

Finalmente, los Wavelet Packets pueden verse como una familia de bases con estructura de árbol, inducidas por el par de filtro de dos canales del tipo CMF $(h(n), g(n))$ como se muestra en la figura 3.1a.



(a) Descomposición es sub-espacios

(b) Implementación mediante banco de filtros

Figura 3.1: Estructura de árbol binario de Wavelet Packets.

Una propiedad clave de las bases asociadas a cada WP es la relación entre los coeficientes transformados, dada por las ecuaciones (3.1) y (3.2), entre distintos niveles de escala. Mas precisamente para $\forall j > L$, sea $x(t) \in U^p_j \subset \mathbb{X}$ con coeficientes

$$d^p_j(n) \equiv \langle x(t), \phi^p_j(t - 2^j n) \rangle, \ \forall n \in \mathbb{Z}. \tag{3.10}$$

La proyección en la base asociada con la descomposición $U^{2p}_{j+1} \bigoplus U^{2p+1}_{j+1}$, se obtiene [24, Proposición 8.4],

$$d^{2p}_{j+1}(n) = \sum_{k\in\mathbb{Z}} d^p_j(k) \cdot h(k-2n), \quad d^{2p+1}_{j+1}(n) = \sum_{k\in\mathbb{Z}} d^p_j(k) \cdot g(k-2n), \quad \forall n \in \mathbb{Z}, \tag{3.11}$$

Y si además se considera que las bases son ortonormales, la *Relación de Parseval* implica

$$||x(t)||^2 = \sum_{n\in\mathbb{Z}} \left|d^p_j(n)\right|^2 = \sum_{n\in\mathbb{Z}} \left|d^{2p}_{j+1}(n)\right|^2 + \sum_{n\in\mathbb{Z}} \left|d^{2p+1}_{j+1}(n)\right|^2. \tag{3.12}$$

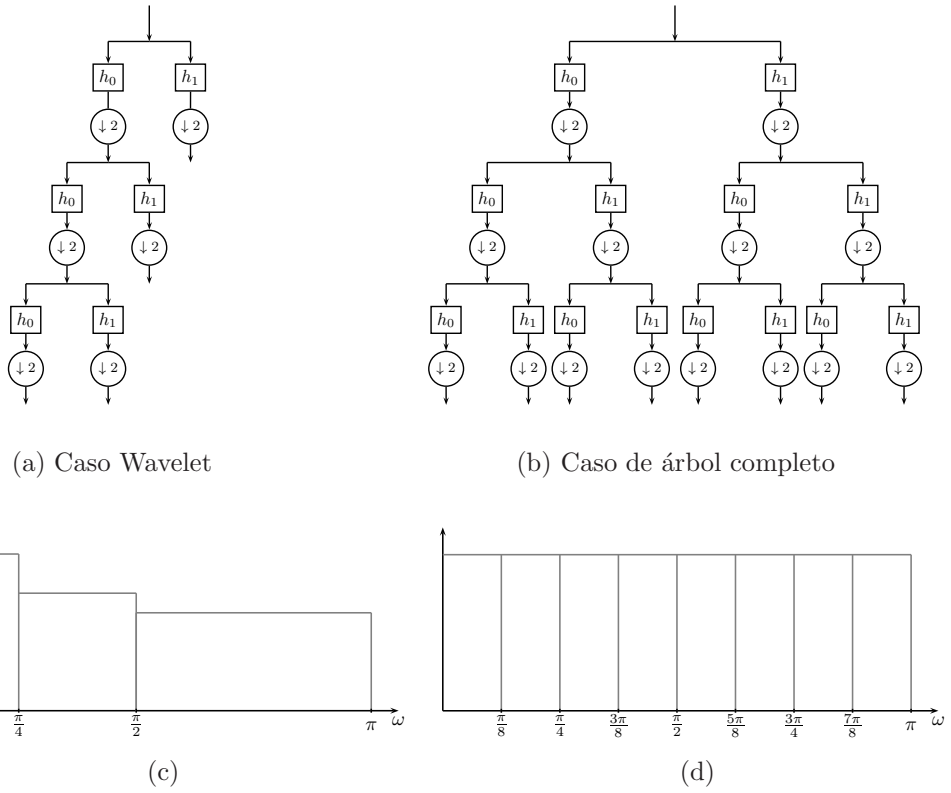
Siguiendo este proceso inductivamente, se puede obtener una fórmula cerrada para los coeficientes transformados, para cada par de bases de una descomposición mediante WP.
La importancia de este resultado es que se logra pasar de un análisis en *tiempo continuo*(3.10), a uno en *tiempo discreto* (algoritmos), en la ecuación (3.11). De hecho asumiendo que $x(t)$ vive en un espacio de resolución finita $\mathbb{X}$, la ecuación (3.10) con $j = L$ y $p = 0$ puede verse como un *Teorema de muestreo generalizado para WP* [26, 27].
Por lo tanto, la estructura de árbol binario de los WP manifestada en la ecuación (3.11) permite una implementación algorítmica rápida de la ecuación de análisis WP (proyección sobre una base)[24]. La siguiente sección profundiza en la implementación de WP mediante bancos de filtros[9].

3.2. Implementación de un WP como banco de filtros multi-tasa

El componente básico utilizado en el análisis mediante WP es el filtro de dos canales, usado para construir distintos árboles de análisis, figuras 3.3 y 3.4. Para la implementación en tiempo discreto, se utiliza el par de CMF como bloque fundamental, la iteración básica se realiza mediante la aplicación de dicho filtro de dos canales, con respuestas al impulso $h(n)$ y $g(n)$, seguido de un operador de sub-muestreo (*downsampler*) por 2 [9, 24].
Esta visión puede ser generalizada para cualquier proyección desde un espacio de resolución finita. Sea $x(t) \in \mathbb{X}$ (nuestro espacio de escala finita 2^L), con coeficientes transformados dados, $(d_L^0(n))_{n\in\mathbb{Z}}$ (obtenidos de ec.(3.10)) que caracterizan completamente a la señal $x(t)$, podemos obtener los coeficientes transformados $(d_j^p(n))_{n\in\mathbb{Z}}$ asociados con el sub-espacio U_j^p, aplicando un filtro discreto seguido de un operador de sub-muestreo de 2^{j-L}, $\forall j > L$ y $\forall p \in \left\{0, .., 2^{j-L} - 1\right\}$ [9]. Este resultado se obtiene directamente de la siguiente propiedad de bancos de filtros multi-tasa.

PROPOSICIÓN 1 *Intercambio de filtrado con sub-muestreo,[9, Cap. 2,pag. 72-73]. Sea $h(n)$ la respuesta al impulso de un sistema LTI con función de transferencia $H(z)$. Para cualquier $(x(n)) \in \mathbb{R}^{\mathbb{Z}}$, es equivalente pasar $x(n)$ por un sub-muestreo de factor N y luego por un sistema LTI con función de transferencia $H(z)$, que pasar $x(n)$ por $H(z^N)$ y luego por un sub-muestreo de factor N. (Figure 3.2).*

Figura 3.2: Intercambio entre sub-muestreo y filtrado



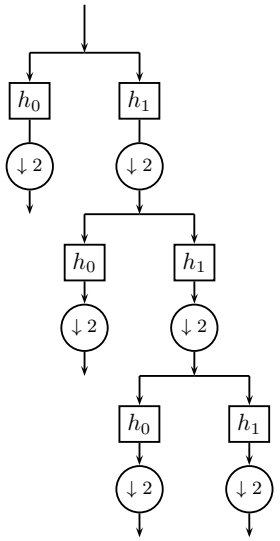
(a) Caso Wavelet

(b) Caso de árbol completo

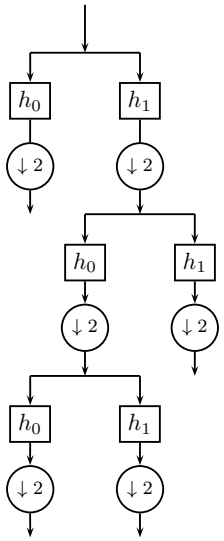
(c)

(d)

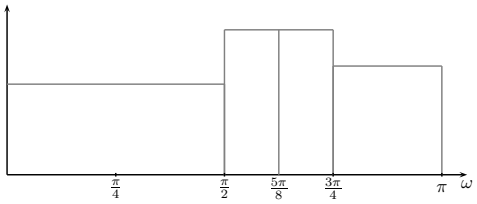
Figura 3.3: Respuesta en frecuencia (ganancia de amplitud) de bancos de filtros equivalentes a un WP. Ejemplo de la partición en frecuencia posible mediante WP para dos estructuras de árbol distintas. Se considera el par de *CMF de Shannon*, que permite una partición perfecta del intervalo $[-\pi, \pi]$. Escenario (3.3a-3.3c) muestra la iteración de $H_0(z)$ (tipo Wavelet), y el escenario (3.3b-3.3d) un árbol completo (resolución uniforme en frecuencia).



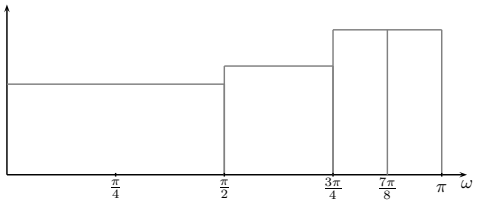
(a) Iteración WP a la derecha



(b) Caso recíproco al wavelet



(c)



(d)

Figura 3.4: Respuesta en frecuencia (ganancia de amplitud) de bancos de filtros equivalentes a un WP. Escenario alternativo al de la figura 3.3. Figuras (3.4a-3.4c) muestran la iteración de $H_1(z)$, y en (3.4b-3.4d) el caso recíproco, en términos de partición en frecuencia, del caso Wavelet de la figura 3.3a.

De la proposición 1, se hace simple derivar una relación entre $(d_L^0(n))_{n\in\mathbb{Z}}$ y los coeficientes transformados de cualquier subespacio U_j^p.

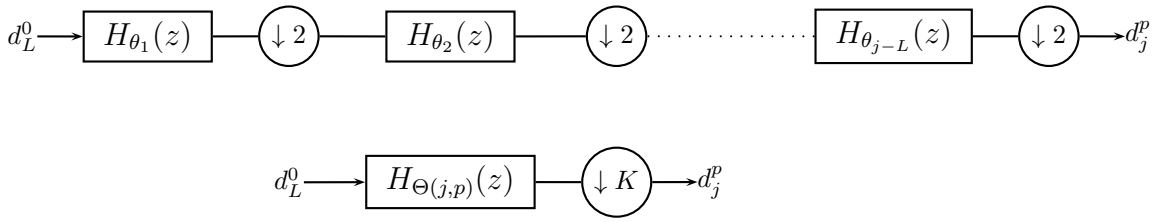


Figura 3.5: Sistemas equivalentes de la proposición 2. El operador de sub-muestreo es de $K = 2^{j-L}$.

PROPOSICIÓN 2 *Filtro Iterado [28, Cap. 11.3.3]. Sea $x(t) \in \mathbb{X}$, un espacio de escala finita 2^L, con coeficientes transformados $(d_L^0(n))_{n\in\mathbb{Z}}$. Si consideramos un sub-espacio arbitrario U_j^p, asociado a una hoja de la descomposición mediante WP, con $j > L$ y $p \in \{0, .., 2^{j-L} - 1\}$. Denotemos por $(h_0(n))_{n\in\mathbb{Z}}$ y $(h_1(n))_{n\in\mathbb{Z}}$, al par de CMF con funciones de transferencia $H_0(z)$ and $H_1(z)$, por $U_{L+1}^{p_1}, .., U_j^{p_{j-L-1}}$ a la secuencia de sub-espacios intermedios utilizados para ir de $\mathbb{X}$ a U_j^p y a $\Theta(j.p) = (\theta_1, .., \theta_{j-L}) \in \{0, 1\}^{j-L}$ el código binario de dicho camino, donde escoger θ_k significa filtrar con $H_{\theta_k}(z)$ y luego aplicar un sub-muestreo por 2 al paso k de la iteración. Entonces $(d_j^p(n))_{n\in\mathbb{Z}}$ se obtiene al pasar $(d_L^0(n))_{n\in\mathbb{Z}}$ por el filtro*

$$H_{\Theta(j,p)}(z) = \prod_{i=1}^{j-L} H_{\theta_k}(z^{2^{i-1}}), \tag{3.13}$$

y luego aplicar un sub-muestreo por 2^{j-L}.

La demostración de la proposición 2 es consecuencia directa de la proposición 1. El sistema equivalente es mostrado en la figura 3.5.

3.3. Respuesta en Frecuencia del banco de filtros de un WP

El sistema que relaciona $(d_L^0(n))_{n\in\mathbb{Z}}$ con $(d_j^p(n))_{n\in\mathbb{Z}}$, proposición 2, es lineal pero no invariante en el tiempo, por el operador de sub-muestreo. Por lo tanto, no es correcto hablar de la respuesta en frecuencia asociada al proceso de proyectar $x(t)$ en el sub-espacio WP U_j^p. Podemos evitar ese fenómeno al considerar solo la parte invariante en el tiempo, es decir, el filtrado (ec. 3.13) y obviar el sub-muestreo.

Como técnica de análisis, proponemos estudiar la respuesta en frecuencia del sistema lineal e invariante en el tiempo (LTI), justo antes de la etapa de sub-muestreo, de esa forma caracterizar el contenido en frecuencia de cada sub-espacio y con ello determinar el tipo de partición en frecuencia asociada a una base WP dada.
Para ilustrar dicho fenómeno, consideremos los *Wavelet Packets de Shannon* [24] inducidos por los filtros ideales pasa-bajos y pasa-altos (ver figuras 3.3 y 3.4),

$$|H_0(e^{j\omega})| = \begin{cases} \sqrt{2} & \omega \in [-\pi/2 + 2k\pi, \pi/2 + 2k\pi], k \in \mathbb{Z} \\ 0 & \text{si no} \end{cases}$$

y

$$|H_1(e^{j\omega})| = \begin{cases} \sqrt{2} & \omega \in [\pi/2 + 2k\pi, 3\pi/2 + 2k\pi], k \in \mathbb{Z} \\ 0 & \text{si no} \end{cases}$$

De acuerdo a lo indicado en la sección 3.1, cada base WP asociada a $\mathbb{X}$ puede representarse por las hojas de un árbol binario (figura 3.1a), es decir, por $\{(j_i, p_i) : i = 1, .., M\}$[1] asociados a los elementos de la base $B = \bigcup_{i=1}^{M} \mathbf{B}_{j_i}^{p_i}$ y la descomposición en sub-espacios $\mathbb{X} = \bigoplus_{i=1}^{M} U_{j_i}^{p_i}$. Para cada hoja (j_i, p_i) de este árbol, se puede obtener su filtro equivalente $H^i(z) \equiv H_{\Theta(j_i,p_i)}(z)$ mediante le ecuación (3.13) y consecuentemente, reducir el estudiar un WP mediante la respuesta en frecuencia de un banco de filtros de M canales (figura 3.6), obviando la etapa de sub-muestreo. Ejemplos de la respuesta en frecuencia antes de la etapa de sub-muestreo se presentan en las figuras 3.3 y 3.4. De ellas se desprende que por una estructura tipo wavelet, iterando $H_0(e^{jw})$ en cada nivel, se obtienen una representación que aumenta la resolución en frecuencia en el rango de baja frecuencia.
En general en cada iteración del filtro de dos canales, se reduce a la mitad el soporte en frecuencia del sub-espacio resultante. Sin embargo si se considera el ordenamiento en frecuencia, las versiones sobre-muestreadas (*up-sampled*) de $H_0(z)$ y $H_1(z)$ no juegan un rol pasa-bajo y pasa-alto respectivamente, en la banda de interés. Esto se debe a que los lóbulos laterales, fuera del rango natural $[-\pi, \pi]$ donde se define la DFT, son traídos al intervalo $[-\pi, \pi]$ de forma no trivial, luego de la operación de sobre-muestreo, ver detalles en [24]. Un ejemplo de este fenómeno se muestra en la figura 3.4a, para el caso de iterar $H_1(z)$ en cada paso, y se muestra que a cada nivel no se aumenta la resolución en la banda de alta frecuencia, como se

[1]Es necesario que $j_i > L$ y $p_i \in 0, .., 2^{j_i - L} - 1$, $\forall i \in \{1, .., M\}$. Además existen condiciones estructurales que garantizan que $\{(j_i, p_i) : i = 1, .., M\}$ corresponden a las hojas de un árbol binario con raíz $(L, 0)$. Se explica con mas detalle en [29, 30, 31] .

podría esperar. Por otro lado la partición en frecuencia recíproca al caso wavelet se muestra en la figura 3.4c, y el modo de iterar el filtro de dos canales no es obvia.

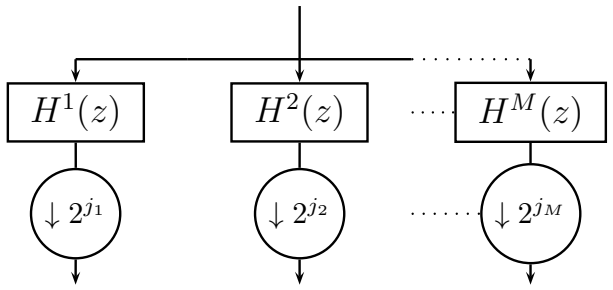


Figura 3.6: Banco de Filtros equivalente de M canales, para una base WP $B = \bigcup_{i=1}^{M} \mathbf{B}_{j_i}^{p_i}$

3.4. Ordenamiento de Frecuencia

Existe una regla simple para re-etiquetar cualquier nodo (j, p) del arbol de un WP, en un nodo equivalente (j, k), a la misma profundidad, de forma que la etiqueta resultante esta ordenada según frecuencia.

PROPOSICIÓN 3 *Código de Gray [24, Cap. 8.1.2]. Sea (j, p) un nodo admisible de la descomposición mediante WP, con camino binario $\Theta(j.p) = (\theta_1, .., \theta_{j-L}) \in \{0, 1\}^{j-L}$, entonces su etiqueta ordenada en frecuencia (j, k) se construye mediante la siguiente regla,*

$$k = G(p) \equiv \sum_{i=1}^{j-L} \bar{\theta}_i \cdot 2^i \in \left\{0, .., 2^{j-L} - 1\right\}, \tag{3.14}$$

Donde $\bar{\theta}_i \equiv \left(\sum_{l=i}^{j-L} \theta_l\right) \ mod\ 2 \in \{0, 1\}$, $\forall i \in \{1, .., j - L\}$.

Al interpretar este resultado para cada base WP $\mathbf{B} = \bigcup_{i=1}^{M} \mathbf{B}_{j_i}^{p_i}$, es posible determinar los índices ordenados $\{(j_i, k_i) : i = 1, .., M\}$, con $k_i = G(p_i)$, (3.14), donde cada sub-espacio $U_{j_i}^{p_i}$, captura la información de la señal concentrada en la banda de frecuencia

$$I_{j_i}^{k_i} \equiv [-k_i \pi 2^{-j_i}, -(k_i + 1)\pi 2^{-j_i}] \cup [k_i \pi 2^{-j_i}, (k_i + 1)\pi 2^{-j_i}]. \tag{3.15}$$

Entonces $\mathbf{B}$ induce $\mathbf{I_B} = \left\{I_{j_i}^{k_i} : i = 1, .., M\right\}$, una partición del rango de frecuencia asociado a la DFT en $[-\pi, \pi]$.

Extendiendo este análisis a un WP con un par de CMF arbitrario $(h_0(n), h_1(n))$, la selectividad en frecuencia dependerá en la concentración de $H_0(e^{jw})$ en el intervalo $[-\pi/2, \pi/2]$. Consecuentemente, solo se tiene una aproximación de la perfecta selectividad en frecuencia alcanzada por los WP de Shannon en (3.15).

Para la aplicación a señales de voz, este aspecto resulta ser crítico. Para ello nos concentramos en la familia de *Wavelet Packets de Daubechies* [24, 7], considerando distintos ordenes para el par de CMF (asociados con el número de ceros en π de $H_0(z)$), lo que conlleva un compromiso entre el orden del filtro de dos canales y la concentración de $H_0(e^{jw})$ en el rango $[-\pi/2, \pi/2]$, o selectividad en frecuencia (detalle en [24, Cap. 8.1.2]).

Capítulo 4

Selección de Banco de Filtros Wavelet Packet

El aspecto central en la implementación del algoritmo de extracción de características propuesto, es la selección del banco de filtros (o selección de una base WP), el cual considera criterios afines a la tarea de reconocimiento fonético a abordar. Se utilizan las metodologías propuestas independientemente por Etemad *et al.* [12] y Saito *et al.* [19][1], y luego re-formulado por Silva *et al.* [2].

El enfoque se basa en la utilización de datos de entrenamiento supervisados para seleccionar una estructura de WP (o una partición en frecuencia de $[-\pi, \pi]$), la que provee una solución cuasi-óptima al problema de discriminación fonética, ver [2, 32, 14]. Antes de formular el problema de optimización, se presenta a continuación una revisión de la notación usada para árboles binarios utilizada en [2].

El proceso de generar una base particular de la familia de WP se representa por un árbol binario [31]. Por simplicidad, sea $J > 0$ el máximo número de iteraciones del proceso de descomposición. Si $G = (V, E)$ es un grafo con $V = \big\{(0,0), (1,0), (1,1), ..., (J,0), ..., (J, 2^J - 1)\big\}$ sus nodos, figura 3.1a, y E la colección de arcos en $V \times V$ que caracterizan un árbol binario completo con nodo raíz $v_{root} = (0,0)$. Se utiliza la convención usada por Breiman *et al.*[29], donde sub-grafos son representados por un subconjunto de nodos del grafo completo. De esta forma, cualquier versión podada del árbol completo representa una base WP particular, o una forma de iterar el filtro de dos canales $(h_0(n), h_1(n))_{n\in\mathbb{Z}}$.

[1]Este trabajo nace motivado del trabajo de Coifman y Wickerhauser [20] en un contexto de selección de bases para representación *sparse* de señales.

Definimos un *árbol binario* $\mathcal{T} = \{v_0, v_1, ..\} \subset V$ como una colección de nodos, uno con grado 2 (nodo raíz) y los de grado 3 (nodos internos) y finalmente nodos terminales (hojas)[33]. Definimos $L(\mathcal{T})$ como el conjunto de hojas de $\mathcal{T}$ y $I(\mathcal{T})$ al conjunto de nodos internos, por lo tanto $L(\mathcal{T}) \cup I(\mathcal{T}) = \mathcal{T}$. Decimos que el árbol binario $\mathcal{S}$ es un sub-árbol de $\mathcal{T}$ si $\mathcal{S} \subset \mathcal{T}$. Además si la raíz de $\mathcal{S}$ y $\mathcal{T}$ son la misma, entonces $\mathcal{S}$ es un sub-árbol podado de $\mathcal{T}$, denotado como $\mathcal{S} \ll \mathcal{T}$. Adicionalmente si la raíz de $\mathcal{S}$ es un nodo interno de $\mathcal{T}$, entonces $\mathcal{S}$ se denomina rama de $\mathcal{T}$. En particular denotamos a la rama mas larga de $\mathcal{T}$ con raíz $v \in \mathcal{T}$ como $\mathcal{T}_v$.
Definimos el tamaño de un árbol $\mathcal{T}$ como su número de hojas, es decir, la cardinalidad de $L(\mathcal{T})$ denotado como $|\mathcal{T}|$. Finalmente, $\mathcal{T}_{full} = V$ denota el árbol binario completo, consecuentemente la colección de bases WP queda indexada por los árboles binarios admisibles, o sea $\{T \in V : T \ll \mathcal{T}_{full}\}$.
Mas precisamente, sea $L(\mathcal{T}) = \{(j_i, p_i) : i \in \{1, .., M\}\}$ un árbol binario WP admisible determinado por sus hojas, entonces su base se denota como $\mathbf{B}_\mathcal{T} \equiv \bigcup_{i=1}^{M} \mathbf{B}_{j_i}^{p_i}$, la descomposición en sub-espacios $\mathbf{U}_\mathcal{T} \equiv \left\{U_{j_i}^{k_i} : i = 1, .., M\right\}$ ($\mathbb{X} = \bigoplus_{i=1}^{M} U_{j_i}^{p_i}$) y la partición ideal en frecuencia $\mathbf{I}_\mathcal{T} \equiv \left\{I_{j_i}^{k_i} : i = 1, .., M\right\}$, con $k_i = G(p_i)$.
La finalidad de este trabajo, es una extensión del análisis cepstral para extracción de características de la sección 2.2, por ello a continuación se muestra la metodología para la extracción de características de energía con WP.
Para cada $\mathcal{T} \ll \mathcal{T}_{full}$, y para alguna señal $x \in \mathbb{X}$, se define el *mapa de energía de banco de filtros* de x relativo a $\mathcal{T}$ como

$$m_\mathcal{T}(x) \equiv \left(E_j^p(x)\right)_{(j,p)\in L(\mathcal{T})} \tag{4.1}$$

Donde $E_j^p(x)$ corresponde a la energía de x en el sub-espacio U_j^p, y por ortonormalidad de la descomposición WP se tiene $||x||_{l_2}^2 = \sum_{(j,p)\in\mathcal{L}(\mathcal{T})} E_j^p(x)$.

4.1. Formulación del problema: Podado del árbol binario

En esta sección se revisita el enfoque de [2], donde la selección de una base WP se basa en el criterio de mínima probabilidad de error de decisión (*Minimun Probability of Error Decision* MPE). Esta formulación reduce el problema a encontrar un balance óptimo entre el error de estimación y aproximación, y por lo tanto tratar con un problema de regularización

de complejidad. En definitiva, la solución de un problema del tipo

$$\mathcal{T}^*(\lambda) = \arg \min_{\mathcal{T} \ll \mathcal{T}_{full}} -F(m_{\mathcal{T}}(X); Y) + \lambda \Phi(\mathcal{T}), \tag{4.2}$$

Donde X es el objeto aleatorio que representa la observación acústica en el espacio de señales $\mathbb{X}$, e Y es la variable aleatoria de la etiqueta, que toma valores en un alfabeto finito de clases fonéticas $\mathbb{Y}$. El primer término en la ecuación (4.2) es $F(\cdot, \cdot)$, una medida de fidelidad que captura la información discriminadora de $m_{\mathcal{T}}(X)$ relativo al problema de clasificación representado por los valores de Y, y el segundo término $\phi(\cdot)$ es una función no-decreciente (término de costo) diseñada para incorporar el efecto del error de estimación (*curse of dimensionality*).
La solución de (4.2) reside en el conjunto solución del siguiente *problema costo-fidelidad*, [31, 2],

$$\mathcal{T}^{k*} = \arg \max_{\{\mathcal{T} \ll \mathcal{T}_{full} : |\mathcal{T}| \leq k\}} F(m_{\mathcal{T}}(X); Y) \tag{4.3}$$

La formulación anterior equivale a encontrar el banco de filtros de tamaño k (o la descomposición WP en k sub-espacios) que maximiza la fidelidad $F(m_{\mathcal{T}}(X); Y)$, para todo $k \in \{2, 3, ..., |\mathcal{T}_{full}|\}$.

Cuando la medida de fidelidad es *aditiva*[2] o alternativamente *afín*[3], con respecto a la estructura de $\mathcal{T}$, que corresponde al tipo de medidas de fidelidad evaluadas en este trabajo. La solución de (4.3) entonces admite una implementación eficiente [2, Theorem 2] y , además tiene una estructura embebida [2, Theorem 3], o sea $\mathcal{T}^{2*} \ll \mathcal{T}^{3*} \ll ... \ll \mathcal{T}^{(|\mathcal{T}_{full}|-1)*} \ll \mathcal{T}_{full}$. El algoritmo para resolver (4.3) se presenta en la sección 4.3.

4.2. Medidas de Fidelidad

Sea $\{(x_i, y_i)\}_{i=1}^N$ realizaciones independientes e idénticamente distribuidas (i.i.d.) del vector conjunto (X, Y), donde cada par (x_i, y_i) corresponde a un segmento de voz y su respectiva etiqueta fonética. Como medidas de fidelidad, se consideran los indicadores propuestos por Saito *et al.* [19], Etermad *et al.* [12] y Silva *et al.*[2]. Todos ellos pueden ser escritos en su

[2]Una función de un árbol $\rho(\cdot)$ es aditivo si $\rho(\mathcal{T}) = \sum_{(j,p) \in L(\mathcal{T})} \rho(j, p)$ [31].

[3]Una función de un árbol $\rho(\cdot)$ se dice afín si, para todo T, S árboles binarios con la misma raíz $S \ll T$, entonces $\rho(T) = \rho(S) + \sum_{s \in \mathcal{L}(S)} \rho(T_s) - \rho(\{s\})$, donde $\{s\}$ es el árbol trivial con raíz s.

forma aditiva:

$$F(m_{\mathcal{T}}(X);Y) = \sum_{(j,p)\in L(\mathcal{T})} F(E_j^p(X);Y). \tag{4.4}$$

El primero es la versión simétrica de la divergencia de Kullback-Leibler (KLD)[34] , propuesta en [19], donde el funcional de hoja corresponde a:

$$F(E_j^p(X);Y) = \sum_{y,z\in\mathbb{Y}} e(j,p,y)\log\left(\frac{e(j,p,y)}{e(j,p,z)}\right), \tag{4.5}$$

y $e(j,p,y)$ denota el mapa de energía, restringido a la clase de etiqueta $y \in \mathbb{Y}$ (ver la definición formal en el Apéndice 7). El segundo indicador es la Información Mutua (IM), adoptada en [2]. Asumiendo la propiedad de árbol markoviano presentada en [2, Prop.3], el funcional es afín [2, Th.1]. Para simplificar la estimación, se asume que las distribuciones condicionadas a la clase fonética son gaussianas, donde la IM reduce a una versión del Indicador Discriminativo de Fisher [32, 35], propuesto por Etermad *et al.* [12]. Mas precisamente:

$$F(E_j^p(X);Y) = tr(S_w^{-1}(t_v)S_b(t_v)) - tr(S_w^{-1}(\{(j,p)\})S_b(\{(j,p)\})). \tag{4.6}$$

En este contexto t_v es un árbol binario con raiz en $v = (j,p)$ y hojas $(j+1,2p)$ y $(j+1,2p+1)$ (ver Figura 3.1a), y $\{(j,p)\}$ es un árbol de un nodo. S_w y S_b corresponden a las matrices de dispersión intra-clase y entre-clases respectivamente. (Detalles en el anexo). Finalmente, como indicador no discriminatorio, se considera la energía promedio por subbanda, propuesta en [18], es decir,

$$F(E_j^p(X);Y) = \frac{1}{N}\sum_{i=1}^{N} E_j^p(x_i). \tag{4.7}$$

Donde considerando la fidelidad de la ecuación (4.7), el algoritmo para resolver (4.3) (sección 4.3) itera en la hoja del árbol $\mathcal{T}^{k*}$ con la mayor energía promedio para encontrar la solución de orden $k+1$, o sea $\mathcal{T}^{k+1*}$.

4.3. Algoritmo de podado

Se presenta una solución al problema de la ecuación (4.3) mediante programación dinámica. El algoritmo utilizado construye una serie de árboles embebidos, y la solución al problema de optimización reside es dicho conjunto. Además esta coincide con la solución del problema

de podado ya que los funcionales utilizados son afines. Detalles de la complejidad computacional y resultados teóricos de este problema se encuentran en [31, 30, 36, 29, 2].
Consideremos el conjunto $(\hat{X}, \hat{Y}) = \{(x_i, y_i)\}_{i=1}^{N}$ de realizaciones i.i.d. de las variables aleatorias (X, Y), donde cada par (x_i, y_i) es un segmento de voz y su respectiva etiqueta fonética, x_i vive en el espacio de resolución finita $\mathbb{X}$ y por simplicidad se asume $L = 0$.

Phase 0: (Choice of parameters)

Choose a specific CMF pair h_0, h_1, a maximum level of decomposition J and a fidelity functional F.

Phase 1: (Computation: Subband measurements and Fidelity Gain)

$\forall j \in \{0, ..., J-1\}, \forall p \in \{0, ..., 2^j - 1\}$ compute:
- $E_j^p(x_i)$: $\forall x_i \in \hat{X}$
$\forall j \in \{0, ..., J-2\}, \forall p \in \{0, ..., 2^j - 1\}$ compute:
- Fidelity gain $\Delta(j, p)$

if (F is KLD functional)
$\Delta(j,p) = F(E_{j+1}^{2p}(\hat{X}); \hat{Y}) + F(E_{j+1}^{2p+1}(\hat{X}); \hat{Y})$
else
$\Delta(j,p) = F(E_j^p(\hat{X}); \hat{Y})$
end

Phase 2: (Initialization)

Initialize: $\mathcal{T}^{2*} = \{(0,0), (1,0), (1,1)\}$, then $L(\mathcal{T}^{2*}) = \{(1,0), (1,1)\}$

Phase 3: (Iteration)

for $k = 2$ to $k = 2^J - 2$

1. - compute: $(j^*, p^*) = \arg \underset{(j,p) \in L(\mathcal{T}^{k*}): j \leq J-1}{\max} \Delta(j,p)$
2. - save: $\mathcal{T}^{(k+1)*} = \mathcal{T}^{k*} \cup \{(j^*+1, 2p^*), (j^*+1, 2p^*+1)\}$

end

Capítulo 5

Soluciones al problema de selección de banco de filtros

5.1. Condiciones Experimentales

La base de datos TIMIT es utilizada en todos los experimentos de este trabajo. TIMIT es una de las bases de datos mas comúnmente usadas para evaluar nuevos métodos y técnicas en Reconocimiento de Voz, principalmente porque es fonéticamente balanceada y cubre un amplio rango de dialectos y personas. Esta variabilidad hace de TIMIT lo suficientemente desafiante para evaluar nuevos métodos en Reconocimiento de Voz, lo que justifica su amplia utilización por la comunidad.

La base de datos consiste de 6300 *frases* de los 8 mayores dialectos de los Estados Unidos. Hay 630 locutores, cada uno diciendo 10 frases. Las transcripciones fonéticas de TIMIT contienen 64 clases fonéticas, y para los experimentos siguientes se ha adoptado el agrupamiento en 39 clases propuesto por [37], además se utiliza un modelo de silencio.

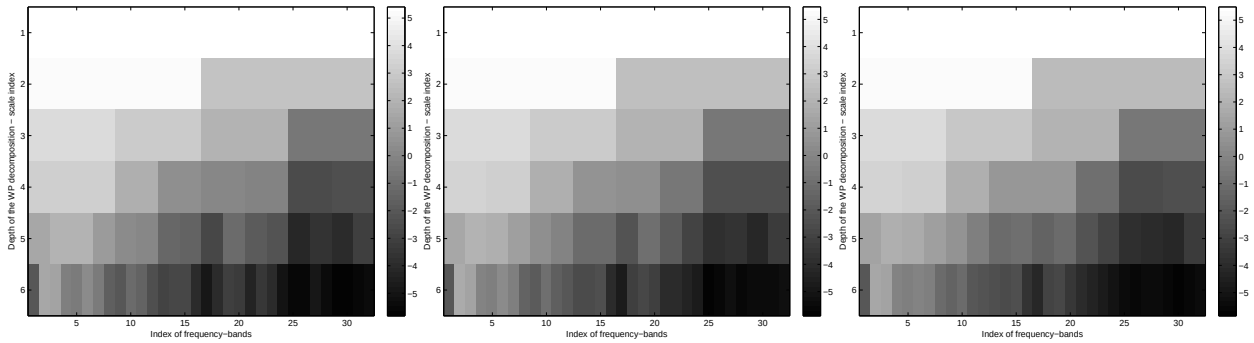
El conjunto de entrenamiento (TRAIN), propuesto en la documentación de TIMIT, fue usado para extraer datos supervisados para la etapa de selección de banco de filtros (Sección 4.1), considerando las segmentaciones temporales y etiquetas fonéticas de las transcripciones de TIMIT.

Para cada señal segmentada, se tomaron 3 trozos de 20ms, de la izquierda, centro y derecha de la señal para asi considerar realizaciones en distintos contextos de los fonemas. Estos se consideraron luego como realizaciones i.i.d. de cada fonema y se utilizaron para calcular las medidas de fidelidad (Fisher, KLD y Energía). Finalmente, con esas medidas de fidelidad se

construyeron los bancos de filtros, soluciones del problema de podado (ec. (4.2)).
Además se han adoptado cuatro diferentes pares de CMF ortogonales, ver Sección 3.2, asociados con los Wavelets de Daubechies de orden 6, 12, 24 y 44 [7, 24, 9]. Con ellos se obtiene una buena cobertura de selectividades en frecuencia, y se agrega una nueva dimensión al análisis, de forma de obtener una mas amplia familia de soluciones al problema de selección de bancos de filtros WP. Es importante mencionar que en este trabajo uno de los aspectos claves es la selectividad en frecuencia, determinada por el orden del par de filtros.

5.2. Análisis de Medidas de Fidelidad

En esta sección se estudia la sensibilidad del algoritmo de selección de banco de filtros (podado de árbol WP) a la selectividad en frecuencia, proporcional al orden de los filtros de Daubechies [24]. Para ello, se analizan las ganancias de fidelidad en función de la resolución en frecuencia (escala j) y localización en frecuencia o banda de frecuencia (asociada al índice de posición k). Se comparan las ganancias de medidas de fidelidad de iterar el filtro de dos canales (ver sección 3.1), para los tres funcionales (Fisher, KLD y Energía).

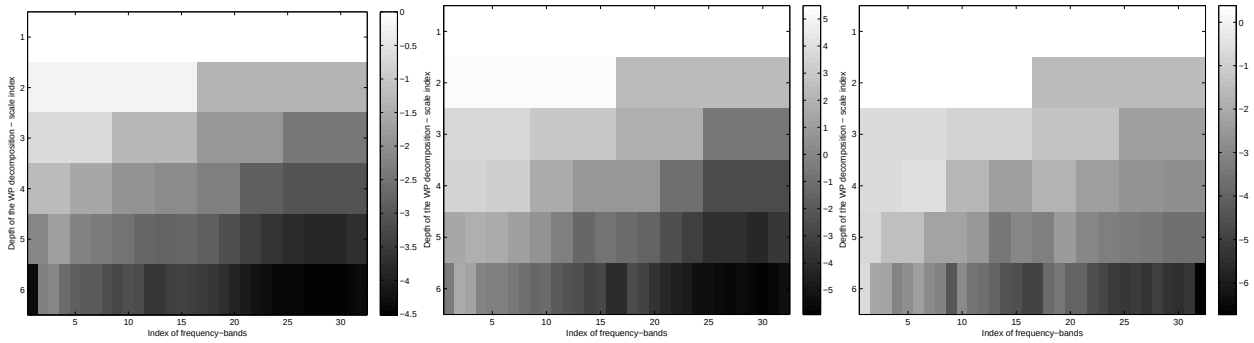


(a) Daubechies orden 6 (b) Daubechies orden 12 (c) Daubechies orden 24

Figura 5.1: Ganancias de funcional KLD indexadas por la escala j (eje vertical) y localización en frecuencia k (eje horizontal), considerando una estructura WP ordenada en frecuencia. Se muestran soluciones para tres selectividades en frecuencia. Colores mas claros indican ganancias mayores.

La figura 5.1 muestra las ganancias de funcional KLD de descomponer el nodo ordenado en frecuencia indexado por (j, k) (asociado con su sub-espacio WP) para los filtros de dos canales Daubechies de orden 6, 12 y 24. Como se espera las mayores ganancias están en el

dominio de las bajas frecuencias, indicando mayor poder discriminatorio de esas bandas.
De la figura, además se observa que las ganancias de funcional KLD mantienen una estructura, y esta evoluciona estabilizándose a medida que aumenta la selectividad en frecuencia (u orden del wavelet). Este fenómeno de estabilización también se observó para las ganancias de los funcionales Fisher y de Energía. Sin embargo cada uno tiene una estructura particular, como se ve en la figura 5.2.
Este fenómeno sugiere que la selectividad en frecuencia no tiene mayor impacto en las medidas de fidelidad, y consecuentemente en las estructuras de los bancos de filtros obtenidos al resolver el problema de podado del arbol WP (ec. (4.3)).



(a) Ganancia de Energía (b) Ganancia de funcional KLD (c) Ganancia de funcional Fisher

Figura 5.2: Ganancias de fidelidad indexadas por escala j (eje vertical) y localización en frecuencia p (eje horizontal), considerando una descomposición WP ordenada en frecuencia. Se presentan las soluciones obtenidas con waveletes Daubechies de orden 44 para los tres funcionales. Colores mas claros indican ganancias mayores.

La figura 5.2 muestra las ganancias de fidelidad calculadas con la mas alta selectividad en frecuencia (Daubechies 44). Los tres funcionales indican que la información del proceso acústico está conentenida en el rango de bajas frecuencias. Por lo tanto las soluciones óptimas entregarán bancos de filtros con estructuras que iteran con mayor prioridad en dicho rango. Además, el método no discriminativo (Energía), de la figura 5.2a, con respecto alos métodos discriminativos, figuras 5.2c y 5.2b, solo tiene diferencias mínimas. Estas pequeñas diferencias son suficientes para establecer que la energía está embebida en las zonas que presentan mayor capacidad de discriminación entre clases fonéticas, pero aún así se pueden establecer diferencias potenciales en las estructuras de banco de filtros que podrían tener significativos efectos en el reconocimiento fonético.

5.3. Respuestas en Frecuencia de los Bancos de Filtros Equivalentes

Con el fin de contrastar los bancos de filtros obtenidos al resolver el problema de podado con diferentes condiciones de selectividad en frecuencia, la figura 5.3 muestra la respuesta en frecuencia de los bancos de filtros equivalentes (Sección 3.3) obtenidos con WP Daubechies de orden 6 y 44, soluciones al podado con el funcional de Fisher.

Ratificando el análisis de la sección anterior, la selectividad en frecuencia no afecta significativamente la estructura del banco de filtros solución, es decir la iteración particular del filtro de dos canales. Esto se observa en los lóbulos principales, los cuales están centrados en las mismas frecuencias para soluciones con el mismo número de bandas. De hecho, las soluciones de tamaño 6 (figuras 5.3a y 5.3b) y tamaño 14 (figuras 5.3e y 5.3f) son representadas por árboles binarios con la misma estructura, aunque el soporte en frecuencia es claramente distinto.

Respecto al soporte en frecuencia, la tendencia es la siguiente. La familia de Wavelets de Daubechies converge al Wavelet de Shannon, a medida que el órden del CMF aumenta [38, 39], luego el soporte en frecuencia del banco de filtros converge a la partición obtenida por los WP de Shannon, ec. (3.15). Además, para un orden de CMF dado, el soporte en frecuencia de un sub-espacio a una profundidad arbitrariamente grande, se vuelve cada vez mas estrecho, siguiendo el soporte de los WP de Shannon, que a la vez converge a una frecuencia puntual [40, Section 3.2] , [41].

En el régimen tratado, de escala finita, a mayor orden de los CMF, el soporte en frecuencia mas se acerca al de Shannon (Sección 3.4). Por lo tanto, al aumenter al orden del par de CMF, las bandas de frecuencia están mas claramente localizadas, y el traslape entre bandas adyacentes se reduce.

Para cada sub-espacio de un árbol WP, dado por su índice ordenado en frecuencia (j, k), el lóbulo principal está concentrado en el intervalo $\mathbf{I}_j^k$ (ver Proposición 3). Sin embargo también hay lóbulos secundarios, que no son necesariamente adyacentes a la banda $\mathbf{I}_j^k$, caracterizando un patrón de interferencia entre bandas muy complejo, y de ganancia significativa particularmente para soluciones con CMF de menor órden, o baja selectividad en frecuencia.

Al interpretar estos resultados, la proyección en el sub-espacio asociado con un nodo (j, k), contiene información principalmente de la banda de Shannon objetivo $\mathbf{I}_j^k$, información de las

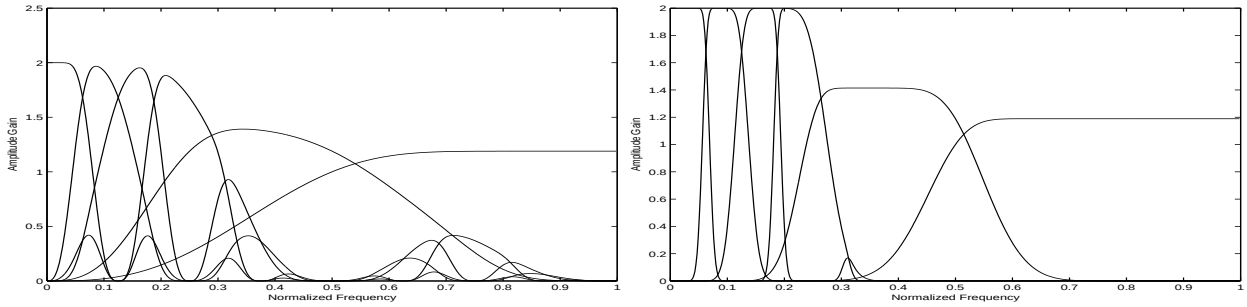
bandas adyacentes, y contenido no despreciables de información de bandas no adyacentes, por las ganancias de lóbulos secundarios, ver figura 5.3.

Afortunadamente, los lóbulos secundarios se desvancen a medida que la selectividad en frecuencia aumenta. Esta tendencia asintótica es justificada formalmente porque la familia de WP Daubechies convergen a los WP de Shannon a medida que aumenta el orden del par de CMF aumenta a infinito [38, 39].

Finalmente en la figura 5.4 se muestra la respuesta en frecuencia de bancos de filtros equivalentes obtenidos mediante metodos discriminativos y no discriminativos. Se usan el WP Daubechies de orden 44, lo que reduce la interferencia de los lóbulos laterales y trae una partición mas clara del intervalo $[-\pi, \pi]$.

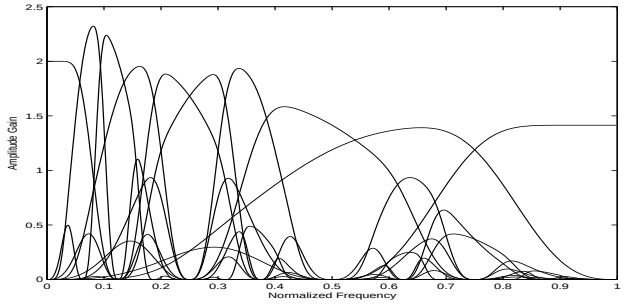
Complementando lo ya observado en las figuras 5.1 y 5.2, las soluciones al problema de podado alcanzan mayores resoluciones en la región de bajas frecuencias. En general, si se comparan los bancos de filtros obtenidos con distintos criterios con el mismo número de bandas, las estructuras son muy similares (figura 5.4), y al aumentar el número de bandas, aparecen diferencias menores.

En conclusión, al estudiar la señal de voz mediante los bancos de filtros obtenidos, se puede decir con seguridad que las estructuras son prácticamente independientes del criterio que considera o no discriminación fonética. Esto ratifica los resultados preliminares obtenidos en [2], donde se afirma que el proceso de producción/precepción de voz es óptimo, en el sentido de que pone energía en las bandas de frecuencia que ofrecen mayor discriminación fonética. Estos resultados se relacionan con el contenido de información de la voz para discriminar fonemas, pero no considera por ejemplo, un escenario ruidoso, o la consideración de contexto fonético donde se podrían esperar otras tendencias.

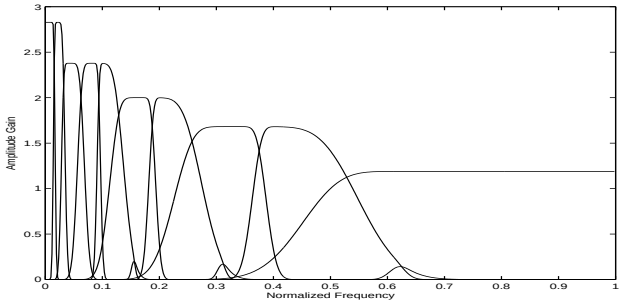


(a) Fisher DB6 6 bandas

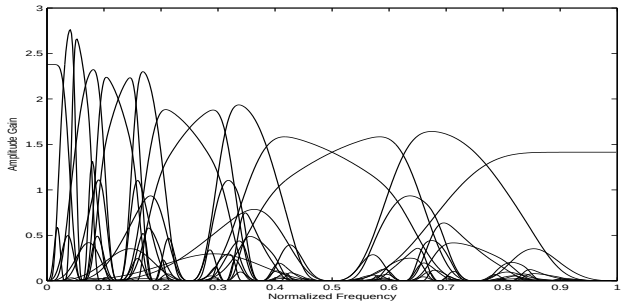
(b) Fisher DB44 6 bandas



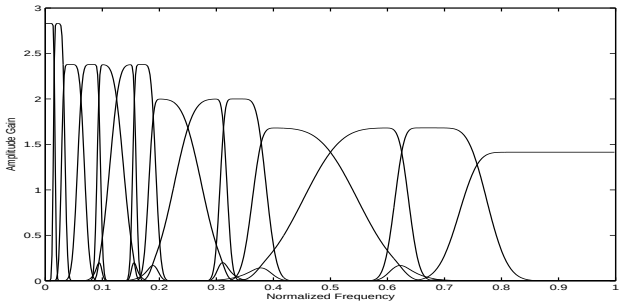
(c) Fisher DB6 10 bandas



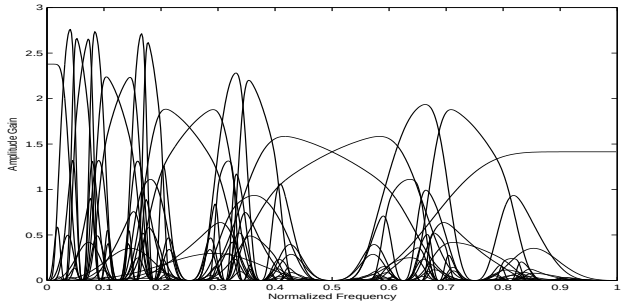
(d) Fisher DB44 10 bandas



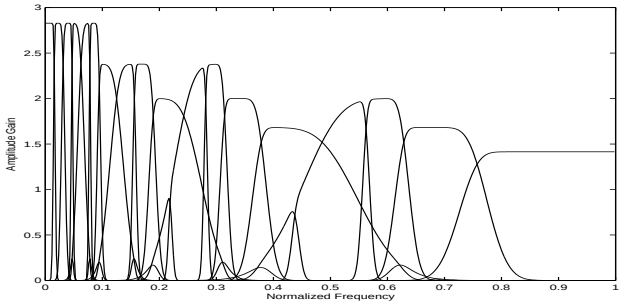
(e) Fisher DB6 14 bandas



(f) Fisher DB44 14 bandas

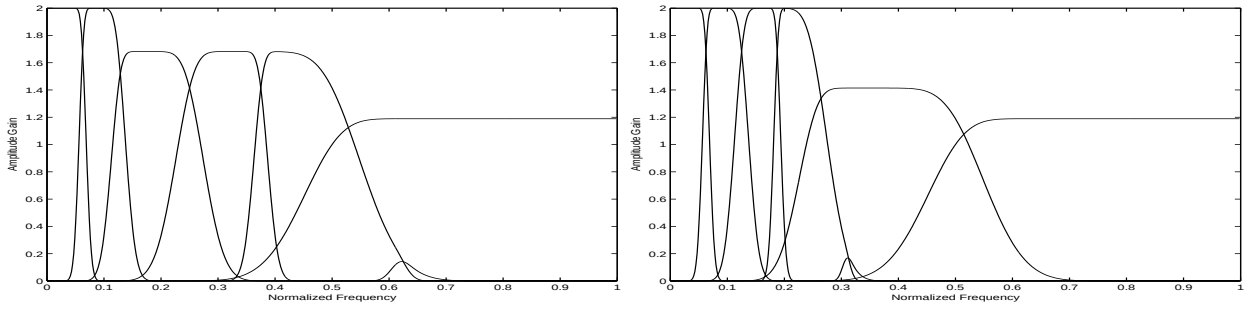


(g) Fisher DB6 18 bandas



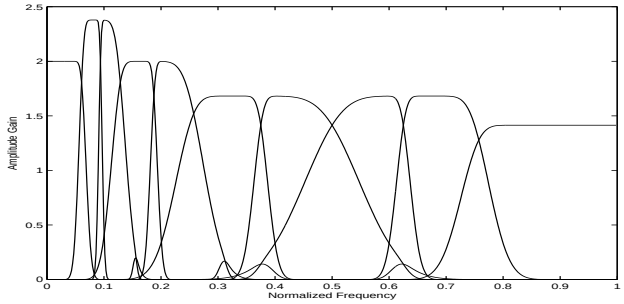
(h) Fisher DB44 18 bandas

Figura 5.3: Respuestas en frecuencia de los bancos de filtros. Soluciones obtenidas con Daubechies de orden 6 (columna izquierda) y 44 (columna derecha), respectivamente.

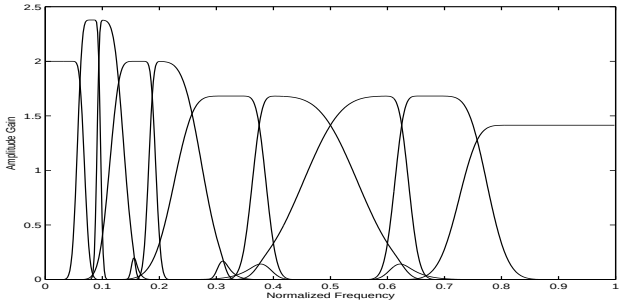


(a) Energy DB44 6 bandas

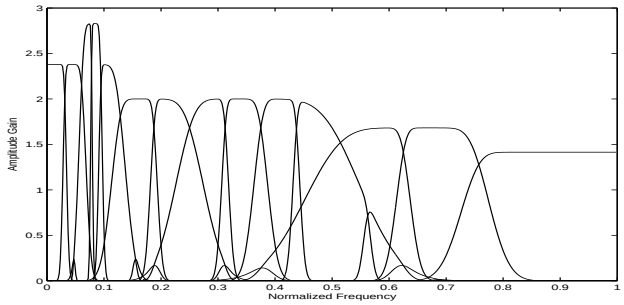
(b) KLD DB44 6 bandas



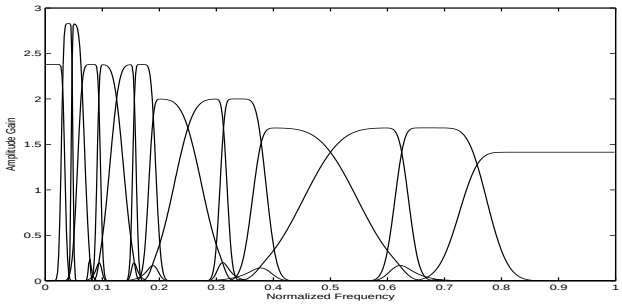
(c) Energy DB44 10 bandas



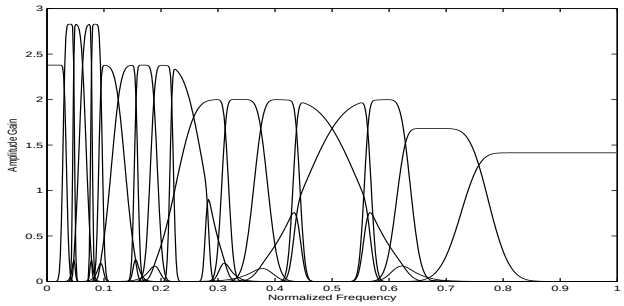
(d) KLD DB44 10 bandas



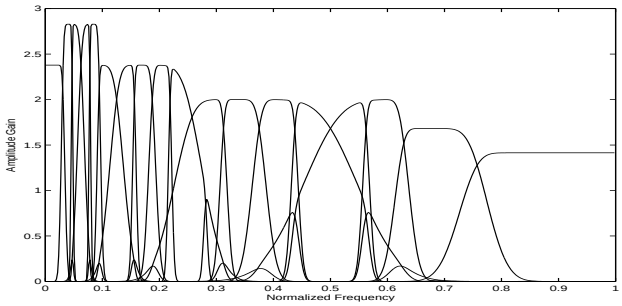
(e) Energy DB44 14 bandas



(f) KLD DB44 14 bandas



(g) Energy DB44 18 bandas



(h) KLD DB44 18 bandas

Figura 5.4: Respuestas en frecuencia de los bancos de filtros. La figura muestra una comparación entre criterios no discriminativo y uno discriminativo, Energia (columna izquierda) y KLD (columna derecha), respectivamente. Todas las figuras corresponden a soluciones obtenidas con Daubechies 44.

Capítulo 6

Experimentos de Reconocimiento Fonético

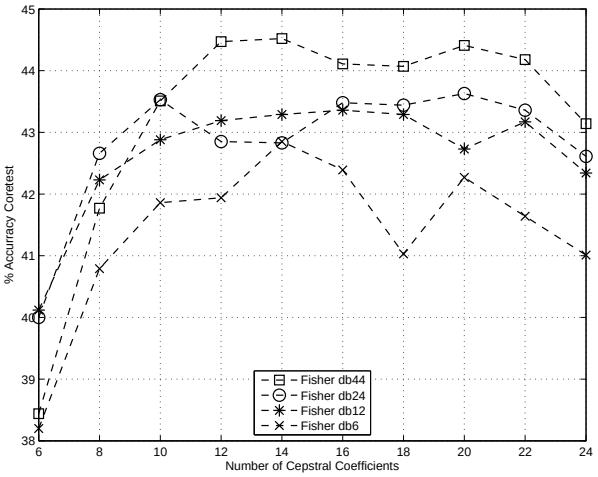
El análisis de esta sección considera varios grados de libertad en el proceso de Extracción de Características mediante WPCC, tales como: la medida de fidelidad para el problema de selección de banco de filtros de la Sección 4.1; la selectividad en frecuencia del filtro de dos canales; el tamaño del banco de filtros; la dimensión del espacio de características.
Como se presentó en secciones anteriores, inducimos los WPCC primero, seleccionando un banco de filtros WP de M-canales; luego derivando los coeficientes de energía ordenados en frecuencia; y finalmente, aplicando la DCT para de-correlación y reducción de dimensión [1] escogiendo los primeros $m < M$ coeficientes de la DCT. Las características WPCC son los resultantes m coeficientes cepstrales mas la log-energía del segmento de voz.
Los experimentos se realizan en una serie de pasos incrementales, con el objetivo de establecer rangos apropiados para todos los grados de libertad mencionados anteriormente. Comenzamos el análisis en un problema de reconocimiento mas simple, de mono-fonemas, que no incluye información contextual en el vector de WPCCs, es decir, no se usan coeficientes delta y aceleración. Esta fase inicial esta diseñada para explorar la dimensión del espacio de características (número de coeficientes cepstrales), y tamaño del WP (número de bandas del banco de filtros), con el objetivo de determinar un rango inicial de valores antes de continuar con experimentos mas complejos. El análisis es efectuado para distintos niveles de selectividad en frecuencia del par de CMF, y para todas las medidas de fidelidad.
Luego se extiende el análisis considerando un vector de características con coeficientes delta y aceleración, bajo el mismo problema de reconocimiento con mono-fonemas, para verificar

si se obtienen las mismas tendencias. Para ello se repiten los experimentos para los rangos de valores considerados en la fase anterior.
Finalmente, se evalúan las características en un experimento de reconocimiento fonético estándar, considerando modelos acústicos contexto-dependientes (tri-fonemas), con un modelo de lenguaje de tipo bi-grama. Como punto de comparación para todos los escenarios, consideramos los MFCC como características estándar, calculadas con un banco de filtros MEL de 22 canales, y seleccionando los 12 primeros coeficientes cepstrales mas log-energía como vector de características. Para cada segmento de voz, se calculan los MFCC o WPCC usando una ventana de Hamming de 32ms, cada 10ms. El sistema de reconocimiento de voz es implementado con el toolbox HTK [22], donde cada modelo acústico es un Hidden Markov Model (HMM) [42], con tres estados emitentes, topología izquierda-derecha, y las observaciones condicionadas a cada estado se modelan con un Gaussian Mixture Model (GMM) de 16 Gaussianas. Se utiliza el esquema de entrenamiento propuesto en la documentación del HTK, y el conjunto Core-test de TIMIT para evaluar el sistema de reconocimiento.

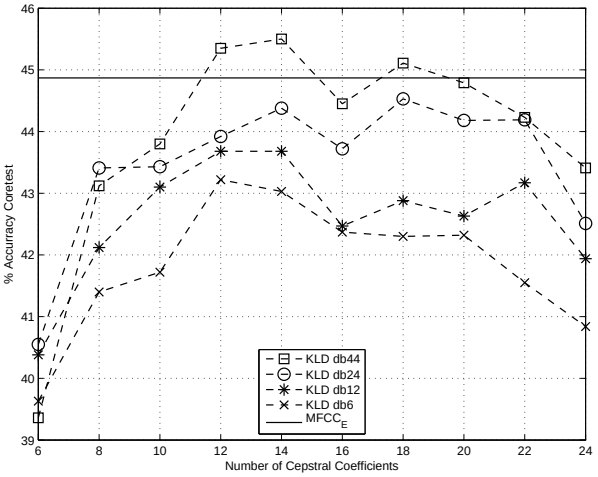
6.1. Experimentos de Reconocimiento fonético contexto-independiente

Se muestran las soluciones de tamaño 24 al problema de podado, para los tres funcionales (KLD, Energia, Fisher). Los vectores de caracteristicas WPCC mas log-energia para bancos de filtros de tamaño fijo, donde se varía el número de coeficientes cepstrales entre 6 y 24, y asi determinar un rango apropiado para la dimension del espacio de caracteristicas. En este contexto, la figura 6.1a muestra las tendencias de reconocimiento fonético medido en *accurracy*[1] para la sulucion obtenida con Fisher a distintas selectividades en frecuencia (CMF Daubechies de orden 6, 12, 24 y 44), variando el número de coeficientes cepstrales.
En cada una de las curvas, se observa el efecto del error de estimación (*curse of dimensionality*). Como se esperaba, hay una tendencia inicial a mejorar el nivel de reconocimiento , para luego producirse una saturación y posterior decrecimiento a medida que aumenta la dimensión del espacio. Este fenómeno se atribuye al fenómeno de error de estimación, bien entendido en problemas de aprendizaje-decisión.

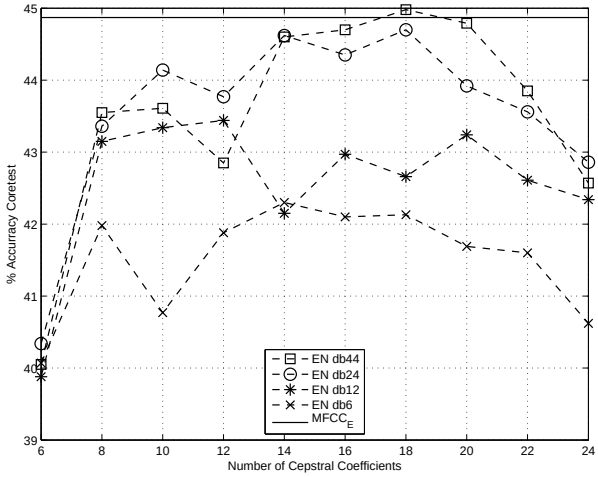
[1] $Accurracy = \frac{N-D-I-S}{N} \times 100\,\%$, donde N es el número total de fonemas en las transcripciones, D, I y S son los errores de eliminación, inserción y sustitución respectivamente, mas detalles ver [22, Capítulo 13.4]



(a) Soluciones Fisher con 24 bandas



(b) Soluciones KLD con 24 bandas

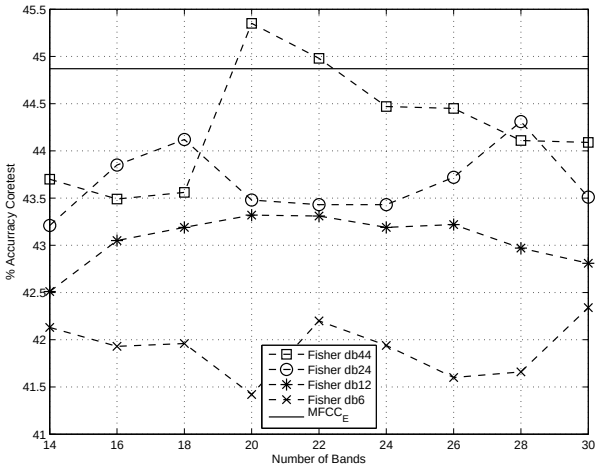


(c) Soluciones Energy con 24 bandas

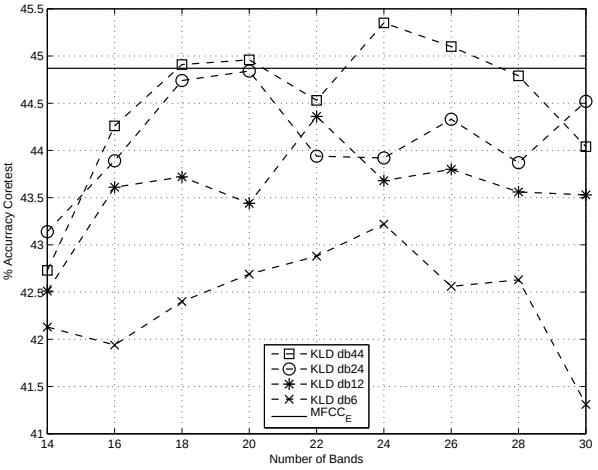
Figura 6.1: *Accurracy* en el conjunto *Core-test* como función del número de coeficientes cepstrales para un tamaño fijo de banco de filtros WP (número de bandas) y características estáticas. Efecto de la selectividad en frecuencia para las soluciones fisher de tamaño 24 (6.1a), soluciones KLD de tamaño 24 (6.1b), y soluciones basadas en la Energía de tamaño 24 (6.1c).

Los resultados muestran un rango óptimo para la dimensión del espacio de características entre 11 y 19. Este rango es prácticamente invariante al considerar otros bancos de filtros óptimos con distinto número de bandas o distintas selectividades en frecuencia. Además este comportamiento es consistente para las otras medidas de fidelidad, KLD y Energía, ejemplificado en las figuras 6.1b y 6.1c, para los WPCC con bancos de filtros de 24 bandas.

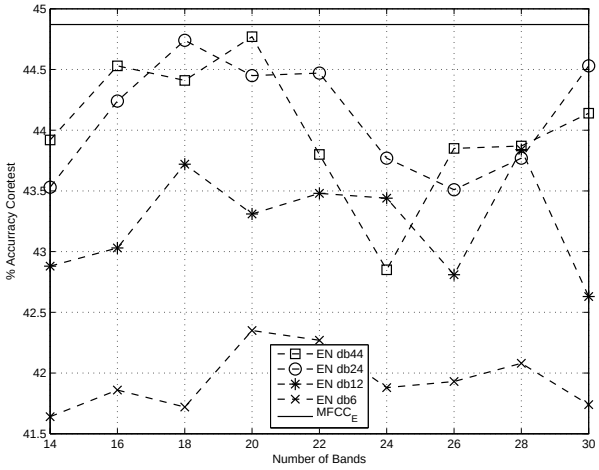
Considerando el *buen* rango de dimensiones para el espacio de características, se fija una de



(a) Fisher, 12 Coeficientes Cepstrales



(b) KLD, 12 Coeficientes Cepstrales



(c) Energy, 12 Coeficientes Cepstrales

Figura 6.2: *Accurracy* en el conjunto *Core-test* como función del tamaño del banco de filtros WP (número de bandas), para 12 Coeficientes Cepstrales y características estáticas. Efecto de la selectividad en frecuencia para bancos de filtro fisher (6.2a), KLD (6.2b) y Energia(6.2c).

ellas, dimensión 13 (con 12 coeficientes cepstrales), para mostrar las tendencias con respecto

al número de bandas de las soluciones. Los experimentos nuevamente consideran todas las medidas de fidelidad y órdenes de los CMF (6, 12, 24 y 44). La figura 6.2 muestra esas tendencias.

Nuevamente se observa un incremento en performance, luego una saturación y finalmente un decrecimiento a medida que aumenta el número de bandas de las soluciones WP. Como en este caso la dimensión del espacio de características esta fijo, no se puede atribuir este efecto al error de estimación (*curse of dimensionality*), sino que al poder de discriminación acústica de los bancos de filtros. De estos resultados se puede inferir que un buen rango para el número de bandas es entre 18 y 26.

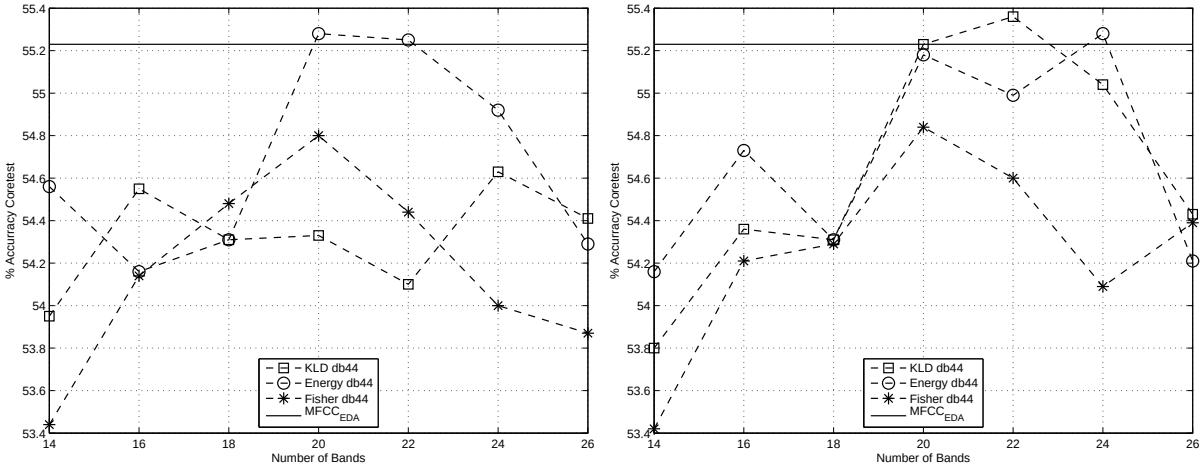
Antes de analizar los resultados de los próximos experimentos, hay que mencionar ciertas tendencias observadas con respecto a la familia de CMF de Daubechies utilizados.

Es interesante observar el efecto de la selectividad en frecuencia (orden del CMF) en los resultados, figuras 6.1 y 6.2. En general, aumentar la selectividad en frecuencia implica mejores porcentajes de reconocimiento fonético, para cualquier dimension, tamaño del banco de filtros y medida de fidelidad. Esto ratifica la conjetura, que la interferencia entre bandas debe ser evitada para la discriminación acústica, y en consecuencia, alcanzar mejores performances al aumentar el orden de los CMF de Daubechies. Estos resultados son consistentes con los reportados en [4], donde mayores selectividades son significativas en tareas de clasificación de segmentos fonéticos.

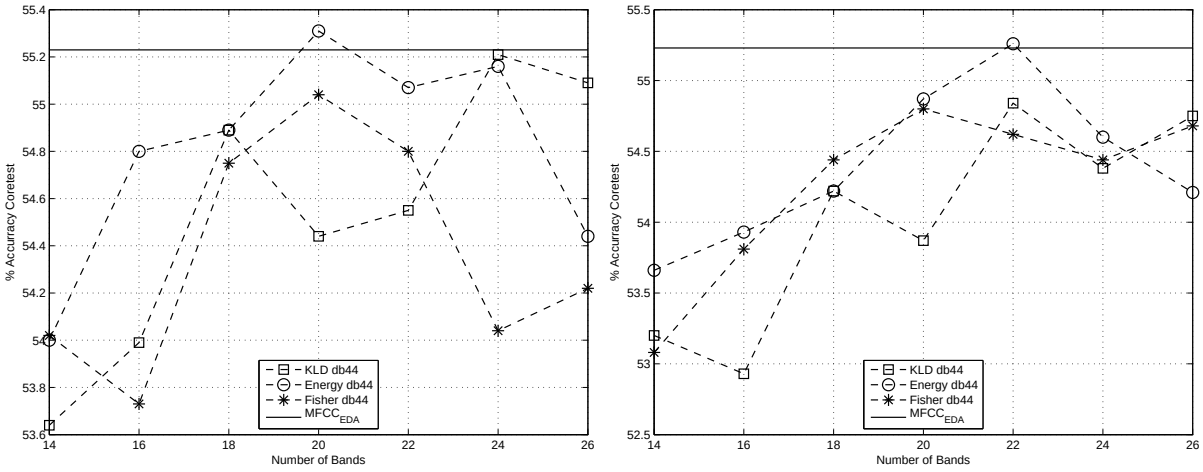
Además es importante notar que se obtuvieron especificaciones concretas para los WPCC que mejoran a los MFCC, en este escenario que no considera información contextual en las características. Los MFCCs alcanzan un porcentaje de *accurracy* de un 44,87 % , de reconocimiento mono-fonético. Se agregan coeficientes delta y aceleración al vector de características. Es un hecho conocido que características dinámicas mejoran el reconocimiento, por lo que se hace interesante estudiar el efecto de estas en las características WPCC. Se consideran escenarios similares a los experimentos anteriores (número de bandas y de coeficientes cepstrales), para explorar el efecto en reconocimiento de las selectividad en frecuencia y de las medidas de fidelidad.

La figura 6.3 muestra el *accurracy* como función del número de bandas, para un número fijo de coeficientes cepstrales en el conjunto $\{11, 12, 13, 14\}$, que corresponde a dimensiones del espacio de caracteristicas $\{36, 39, 42, 45\}$, respectivamente, ademas con la máxima selectividad en frecuencia considerada (DB44).

En general los mejores resultados se obtienen en el rango de 20 a 26 bandas (curvas de

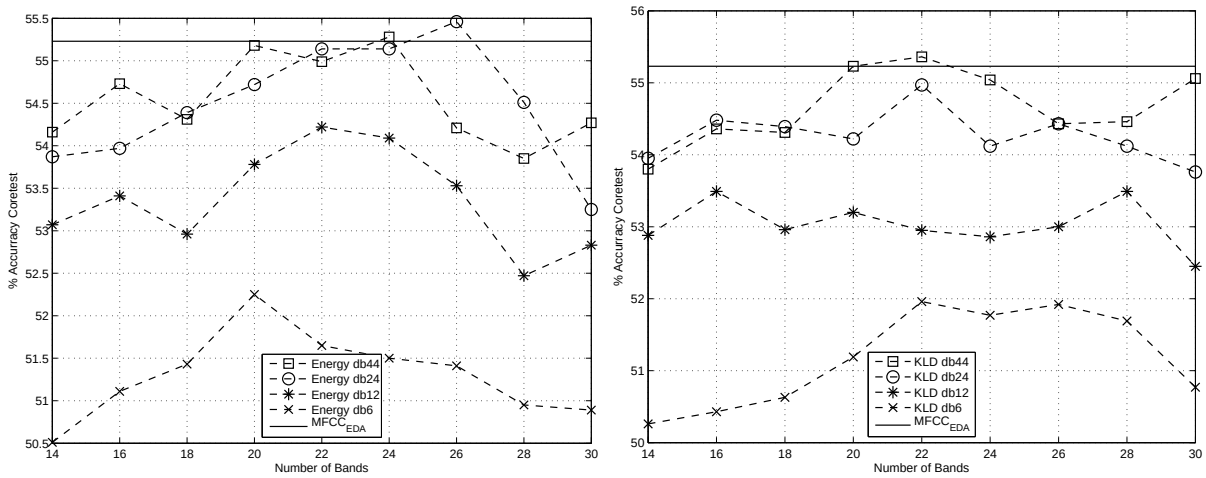


(a) Soluciones DB44 con 11 Coeficientes Cepstrales (b) Soluciones DB44 con 12 Coeficientes Cepstrales



(c) Soluciones DB44 con 13 Coeficientes Cepstrales (d) Soluciones DB44 con 14 Coeficientes Cepstrales

Figura 6.3: *Accurracy* en el conjunto *Core-test* como función del tamaño del banco de filtros WP (número de bandas), para un número fijo de Coeficientes Cepstrales y delta y aceleración. Comparación de métodos, mediante soluciones obtenidas con la mayor selectividad en frecuencia (DB 44).



(a) Soluciones de Energía con 12 Coeficientes Cepstrales

(b) Soluciones KLD con 12 Coeficientes Cepstrales

Figura 6.4: *Accurracy* en el *core-test* como función del tamaño del banco de filtros (número de bandas), para un número fijo de 12 coeficientes cepstrales, y características delta y aceleración. Se comparan soluciones a diferentes selectividades en frecuencia para método de Energía (6.4a) y KLD (6.4b).

figura 6.3). Adicionalmente, el criterio de Energía muestra sistematicamente mejores resultados con respecto a los MFCCs (vector de dimensión 39), con *accurracy* de 55.3 %. A pesar de eso, el mejor resultado particular (marginalmente mejor) se obtiene con la solución de 22 bandas del criterio basado en la KLD, con 12 coeficientes cepstrales (vector de dimensión 39), mostrado en la figura 6.3b.

Por otra parte, la figura 6.4, revisita el efecto de selectividad en frecuencia en el reconocimiento fonético paras las soluciones obtenidas con los métodos de Energía y KLD, con 12 coeficientes cepstrales. Se ratifica que los filtros con mayor orden obtienen mejores resultados. Finalmente se muestra en la tabla 6.1, las ganancias porcentuales al incorporar coeficientes delta y aceleración. Las ganancias aumentan cuando se consideran filtros con mejores selectividades en frecuencia, reafirmando su importancia como criterio de selección de características apropiadas para el reconocimiento fonético.

	DB6	DB12	DB24	DB44
$WPCC_E$	42,09 %	43,29 %	43,97 %	44,26 %
$WPCC_{EDA}$	51,19 %	53,13 %	54,2 %	54,5 %
Ganancia	9,1 %	9,84 %	10,22 %	10,24 %

Tabla 6.1: Ganancias promedio en *Accurracy*, al pasar de características estáticas $WPCC_E$ a dinámicas $WPCC_{EDA}$. Resultados obtenidos en un escenario con 12 coeficientes cepstrales mas log-energía y número de bandas de 14 a 30. La primera fila muestra el reconocimiento medio con características estáticas para los cuatro escenarios de selectividad en frecuencia y promediado entre todos los métodos (KLD, Fisher, Energía). La segunda fila muestra el Accurracy obtenido al correr los mismos experimentos al agregar características dinámicas (delta y aceleración). La tercera columna muestra las ganancias.

6.2. Experimentos de Reconocimiento fonético contexto-dependiente

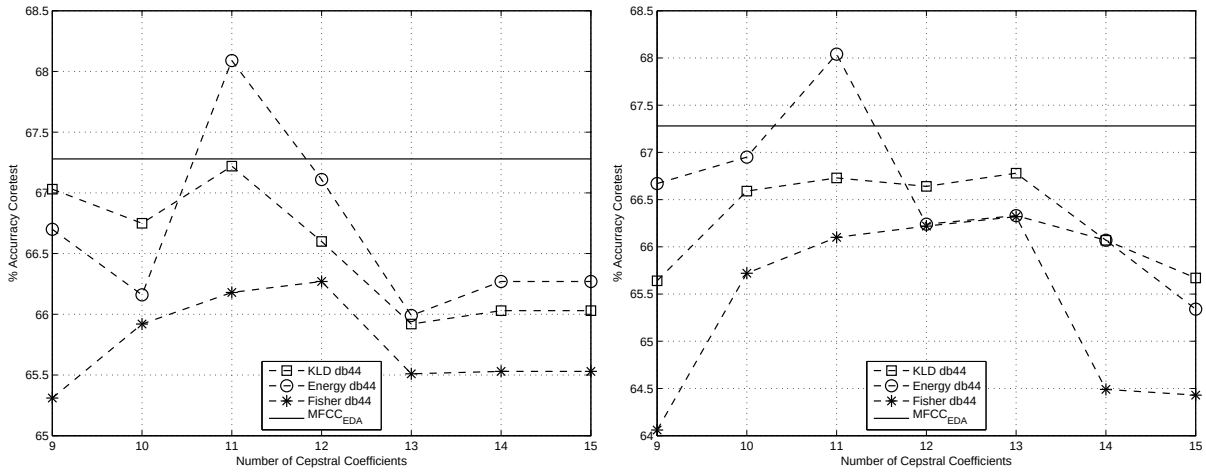
En esta sección se evalua el reconocimiento en una tarea de reconocimiento estandar, que considera HMMs contexto-dependiente, características con coeficientes delta y aceleración y un modelo de lenguage bi-grama. Para esto nos centramos en el análisis del rango de 20 a 26 bandas, y cantidad de coeficientes cepstrales en torno a 12. Estos son los rangos donde se obtuvieron mejores resultados en los experimentos anteriores.

La figura 6.5 muestra el reconocimiento en *accurracy* como función del número de coeficientes cepstrales, donde solo se muestran las mejores tendencias, observadas para soluciones de 24 y 26 bandas con filtros wavelet Daubechies de orden 44. Se considera un número de coeficientes cepstrales entre 9 y 15 (espacio de características entre 30 y 48). El efecto de compensación entre los errores de estimación y aproximación se observa nuevamente, sin embargo es diferente al caso de modelos contexto-independiente, de la figura 6.1. La razón es la siguiente, en este caso, el número de modelos es mayor, asi mismo el número de parámetros a estimar, pero los datos de entrenamiento son los mismos. Entonces el error de estimación domina antes al error de aproximación, para dimensiones del espacio de características mas pequeñas, con respecto a los resultados mostrados en la figura 6.1.

También se observa que los metodos basados en Energía y KLD alcanzan las mejores tenden-

cias, consistente con los experimentos de la sección anterior. En el escenario con 26 bandas y 11 coeficientes cepstrales, y 24 bandas y 11 coeficientes cepstrales, las soluciones basadas en Energía alcanzan *accurracy* de 68.04 % y 68.09 % respectivamente. Resultados competitivos con los MFCC que alcanzan un 67.28 %. Para terminar, los bancos de filtros equivalentes para las soluciones basadas en Energía, con 24 y 26 bandas se ilustran en las figuras 6.6a y 6.6b, respectivamente.

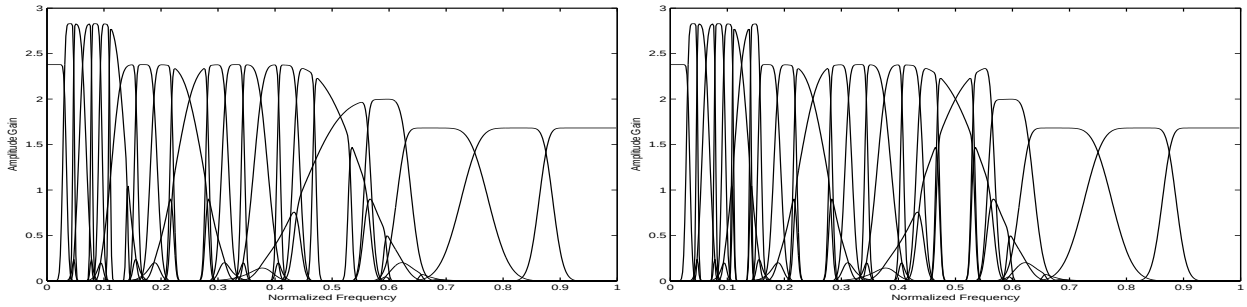
Es interesante notar, que la escala Mel, tiene aproximadamente una partición uniforme en las bajas frecuencias y para el resto del espectro una particion uniforme en el dominio logaritmico [1]. Los bancos de filtros obtenidos de las figuras 6.6a y 6.6b, y desglosados en la tabla 6.2, ofrecen una partición aproximadamente uniforme (con el mismo ancho de banda) en el intervalo $[0, 1KHz]$, y luego anchos de banda crecientes entre $2Hz$ y $8KHz$. Se hace incapié en que el método de este trabajo es dependiente de los datos, y que sigue una partición en frecuencia con estructura muy similar a la escala Mel.



(a) Soluciones DB44 con 24 bandas

(b) Soluciones DB44 con 26 bandas

Figura 6.5: *Accurracy* de reconocimiento fonético con modelos contexto-dependiente en función del numero de coeficientes cepstrales, considerando características delta y aceleración.



(a) Energía DB44 24 bandas (b) Energía DB44 26 bandas

Figura 6.6: Respuestas en frecuencia de los bancos de filtros WP que alcanzan mejores resultados de reconocimiento. Intervalo de frecuencia normalizado sobre $[0, 8KHz]$.

Hoja (j,k)	$(5,0)$	$(6,2)$	$(6,3)$	$(6,4)$	$(6,5)$	$(6,6)$
Banda I_j^k [Hz]	$[0,250]$	$[250,375]$	$[375,500]$	$[500,625]$	$[625,750]$	$[750,875]$
Ancho de banda [Hz]	250	125	125	125	125	125
Hoja (j,k)	$(6,7)$	$(5,4)$	$(5,5)$	$(5,6)$	$(5,7)$	$(5,8)$
Banda I_j^k [Hz]	$[875,1000]$	$[1000,1250]$	$[1250,1500]$	$[1500,1750]$	$[1750,2000]$	$[2000,2250]$
Ancho de banda [Hz]	125	250	250	250	250	250
Hoja (j,k)	$(5,9)$	$(5,10)$	$(5,11)$	$(5,12)$	$(5,13)$	$(5,14)$
Banda I_j^k [Hz]	$[2250,2500]$	$[2500,2750]$	$[2750,3000]$	$[3000,3250]$	$[3250,3500]$	$[3500,3750]$
Ancho de banda [Hz]	250	250	250	250	250	250
Hoja (j,k)	$(5,15)$	$(4,8)$	$(4,9)$	$(3,5)$	$(3,6)$	$(3,7)$
Banda I_j^k [Hz]	$[3750,4000]$	$[4000,4500]$	$[4500,5000]$	$[5000,6000]$	$[6000,7000]$	$[7000,8000]$
Ancho de banda [Hz]	250	500	500	1000	1000	1000

Tabla 6.2: Partición ideal en frecuencia para el intervalo $[0, 8KHz]$, con la solución de la figura 6.6a. Contiene las hojas ordenadas en frecuencia, i.e. $\{(j, k = G(p)) : (j,p) \in \mathcal{L}(T)\}$, sus respectivos soportes (I_j^k) y anchos de banda .

Capítulo 7

Conclusión

Utilizando la amplia familia de bases Wavelet Packet y estudiando las respuestas en frecuencia de sus bancos de filtros equivalentes, se ha analizado la estructura y ubicación de la energía en distintas frecuencias, del proceso de producción de voz. La metodología aplicada permite obtener bancos de filtros WP con estructuras que dan mayor resolución a la región de bajas frecuencias, similar a la escala Mel. Estas estructuras fueron obtenidas mediante criterios discriminativos y no-discriminativos, confirmando la hipótesis que el mecanismo de producción y percepción de voz es un proceso óptimo en poner y distinguir energía e información en el rango de bajas frecuencias.

En reconocimiento de voz, los experimentos muestran las representaciones posibles con un método de extracción de características basado en WP, donde se obtienen niveles de reconocimiento competitivos respecto a los MFCC, sin considerar optimizaciones en el método de entrenamiento. Aún es un tema abierto, si las características WP pueden ser una mejor opción con un estudio mas profundo en las técnicas de entrenamiento y reconocimiento, y si por ejemplo entrenamiento discriminativo puede beneficiarse de la naturaleza de esta metodología de extracción de características, o si la selección de banco de filtros puede aplicarse a otro problema de clasificación, tal como reconocimiento de grupos fonéticos o reconocimiento de locutor.

Se alcanza mayor calidad en las características WP cuando se considera la alta selectividad en frecuencia como un requisito para alcanzar niveles de reconocimiento competitivos. El diseño de wavelets o equivalentemente CMF considerando selectividad en frecuencia también es una linea interesante de seguir, asi como la relación entre selectividad en frecuencia, calidad del espacio de características y las propiedades de decorrelación de wavelets y wavelet packets

para procesos estacionarios.

Anexos

Estimación de Medidas de Fidelidad

Presentamos detalles sobre el calculo de las medidas de fidelidad KLD y Fisher. Para ello, sea $(\hat{X}, \hat{Y}) = \{(x_i, y_i)\}_{i=1}^{N}$ los datos supervisados, o sea, realizaciones i.i.d. del vector aleatorio conjunto (X, Y).

Estimación de la medida de fidelidad KLD

Primero se define la energia normalizada de $x \in \mathbb{X}$ por $\bar{E}_j^p(x) \equiv \frac{E_j^p(x)}{\|x\|^2}$, y el numero de ejemplos en la clase $y \in \mathbb{Y}$ como $N_y \equiv \sum_{i=1}^{N} \mathbb{I}_{\{y\}}(y_i)$. El mapa de energia $e(j, p, y)$ esta dado por

$$e(j, p, y) = \frac{1}{N_y} \sum_{i=1}^{N} \mathbb{I}_{\{y\}}(y_i) \cdot \bar{E}_j^p(x_i), \tag{7.1}$$

para cada par $(j, p) \in \{0, ..., J\} \times \{0, ..., 2^j - 1\}$ y $y \in \mathbb{Y}$. Sea el arbol binario $\mathcal{T}$, su mapa de energia condicionada a una clase se define como

$$e_{\mathcal{T}}(y) = (e(j, p, y))_{(j,p) \in L(\mathcal{T})}, \tag{7.2}$$

donde por la *Relación de Parseval*, tenemos que $\sum_{(j,p) \in L(\mathcal{T})} e(j, p, y) = 1$. Por lo tanto podemos tratar a $e_{\mathcal{T}}(y)$ como una funcion de masa de probabilidad, y definir la fidelidad basada en la KLD como[19]

$$F(m_{\mathcal{T}}(\hat{X}; \hat{Y})) = \sum_{y,z \in \mathbb{Y}} \mathcal{D}(e_{\mathcal{T}}(y) \| e_{\mathcal{T}}(z)). \tag{7.3}$$

El funcional $\mathcal{D}$ es la versión discreta de la KLD [43, 44]. Para escribir el funcional en su forma aditiva, de (4.5), se consideran las siguientes igualdades:

$$\begin{aligned} F(m_{\mathcal{T}}(\hat{X};\hat{Y})) &= \sum_{y,z\in\mathbb{Y}} \mathcal{D}(e_{\mathcal{T}}(y)\|e_{\mathcal{T}}(z)) \\ &= \sum_{y,z\in\mathbb{Y}} \sum_{(j,p)\in L(\mathcal{T})} e(j,p,y)\log\left(\frac{e(j,p,y)}{e(j,p,z)}\right) \\ &= \sum_{(j,p)\in L(\mathcal{T})} \sum_{y,z\in\mathbb{Y}} e(j,p,y)\log\left(\frac{e(j,p,y)}{e(j,p,z)}\right) \\ &= \sum_{(j,p)\in L(\mathcal{T})} F(E_j^p(\hat{X});\hat{Y}). \end{aligned}$$

Estimación de fidelidad Fisher

Sea el vector de energía de una señal x_i en el arbol $\mathcal{T}$ dado por $m_{\mathcal{T}}(x_i) = \left(E_j^p(x_i)\right)_{(j,p)\in\mathcal{T}}$,y $\hat{P}(\{y\}) = \frac{N_y}{N}$ denota la funcion de masa de probabilidad condicionada a la clase $y \in \mathbb{Y}$. Asumiento que la probabilidad condicionada a la clase y, la variable aleatoria $m_{\mathcal{T}}(X)$ es gaussiana, el estimador de maxima verosimilitud de su media y covarianza son:

$$\hat{\mu}_y = \frac{1}{N_y}\sum_{i=1}^{N} \mathbb{I}_{\{y\}}(y_i) m_{\mathcal{T}}(x_i) \tag{7.4}$$

y

$$\Sigma_y = \frac{1}{N_y}\sum_{i=1}^{N} \mathbb{I}_{\{y\}}(y_i)(m_{\mathcal{T}}(x_i) - \hat{\mu}_y)(m_{\mathcal{T}}(x_i) - \hat{\mu}_y)^\dagger, \tag{7.5}$$

respectivamente. El estimador de la media no condicional es $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} m_{\mathcal{T}}(x_i)$.
Ahora se puede definir la matriz de dispersión intra-clase S_w para el árbol $\mathcal{T}$ como

$$S_w(\mathcal{T}) = \sum_{y\in\mathbb{Y}} \hat{P}(\{y\})\cdot\Sigma_y, \tag{7.6}$$

Y la matriz de dispersión entre-clases

$$S_b(\mathcal{T}) = \sum_{y\in\mathbb{Y}} \hat{P}(\{y\})\cdot(\hat{\mu} - \hat{\mu}_y)(\hat{\mu} - \hat{\mu}_y)^\dagger. \tag{7.7}$$

Finalmente para el árbol binario $\mathcal{T}$, el funcional asociado a su hoja (j,p) se define como

$$F(E_j^p(\hat{X});\hat{Y}) = tr(S_w^{-1}(t_v)S_b(t_v)) - tr(S_w^{-1}(\{(j,p)\})S_b(\{(j,p)\})) \tag{7.8}$$

Donde se recupera el funcional asociado al árbol de la ecuación (4.4).

Bibliografía

[1] T. F. Quatieri, *Discrete-time Speech Signal Processing principles and practice*. Prentice Hall, 2002.

[2] J. Silva and S. Narayanan, "Discriminative wavelet packet filter bank selection for pattern recognition," *IEEE Transactions on Signal Processing*, vol. 57, no. 5, pp. 1796–1810, 2009.

[3] O. Farooq and S. Datta, "Mel filter-like admissible wavelet packet structurefor speech recognition," *IEEE Signal Processing Letters*, vol. 8, no. 7, pp. 196–198, 2001.

[4] G. Choueiter and J. Glass, "An implementation of rational wavelets and filter design for phonetic classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 939–948, 2007.

[5] K. Kim, D. Youn, and C. Lee, "Evaluation of wavelet filters for speech recognition," in *IEEE Int. Conf. Syst. Man. Cybern.*, 2000, pp. 2891–2894.

[6] B. Tan, F. Minyue, A. Spray, and P. Dermody, "The use of wavelet transform in phoneme recognition," in *Int. Conf. Spoken Lang. Process.*, 1996, pp. 2431–2434.

[7] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia: SIAM, 1992.

[8] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, July 1989.

[9] M. Vetterli and J. Kovacevic, *Wavelet and Subband Coding*. Englewood Cliffs, NY: Prentice-Hall, 1995.

[10] R. Coifman, Y. Meyer, S. Quake, and V. Wickerhauser, "Signal processing and compression with wavelet packets," Numerical Algorithms Research Group, New Haven, CT, Yale University, Tech. Rep., 1990.

[11] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 46, no. 46, pp. 886–902, April 1998.

[12] K. Etemad and R. Chellapa, "Separability-based multiscale basis selection and feature extraction for signal and image classification," *IEEE Transactions on Image Processing*, vol. 7, no. 10, pp. 1453–1465, October 1998.

[13] K. Ramchandran, M. Vetterli, and C. Herley, "Wavelet, subband coding, and best bases," *Proceedings of the IEEE*, vol. 84, no. 4, pp. 541–560, April 1996.

[14] N. Vasconcelos, "Minimum probability of error image retrieval," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2322–2336, 2004.

[15] A. S. Willsky, "Multiresolution Markov models for signal and image processing," *Proceedings of the IEEE*, vol. 90, no. 8, pp. 1396–1458, August 2002.

[16] R. E. Learned, W. C. Karl, and A. S. Willsky, "Wavelet packet based transient signal classification," 1992, pp. 109 – 112.

[17] C. Scott and R. D. Nowak, "Templar: A wavelet-based framework for pattern learning and analysis," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2264–2274, August 2004.

[18] T. Chang and C. J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Transactions on Image Processing*, vol. 2, no. 4, pp. 429–441, 1993.

[19] N. Saito and R. R. Coifman, "Local discriminant basis," *in Proc. SPIE 2303, Mathematical Imaging: Wavelet Applications in Signal and Image Processing*, pp. 2–14, 1994.

[20] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithm for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713–718, March 1992.

[21] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1983.

[22] S. Young, *The HTK Book (for HTK Version 3.4)*, 2009.

[23] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[24] S. Mallat, *A Wavelet Tour of Signal Processing*, 3rd ed. Academic Press, 2009.

[25] R. R. Coifman, Y. Meyer, and M. V. Wickerhauser, "Wavelet analysis and signal processing," in *Wavelets and their Applications*, B. Ruskai, Ed. Jones and Barlettt, 1992, pp. 153–178.

[26] X. Zhou and W. Sun, "On the sampling theorem for wavelet subspaces," *The Journal of Fourier Analysis and Applications*, vol. 5, no. 4, pp. 347–354, 1999.

[27] G. G. Walter, "A sampling theorem for wavelet subspaces," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 881–884, 1992.

[28] P. P. Vaidyanathan, *Multirate Systems and Filter Banks.* Englewood Cliffs:NY Prentice-Hall, 1993.

[29] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees.* Belmont, CA: Wadsworth, 1984.

[30] P. Chou, T. Lookabaugh, and R. Gray, "Optimal pruning with applications to tree-structure source coding and modeling," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 299–315, 1989.

[31] C. Scott, "Tree pruning with subadditive penalties," *IEEE Transactions on Signal Processing*, vol. 53, no. 12, pp. 4518–4525, 2005.

[32] J. Silva and S. Narayanan, "Minimum probability of error signal representation," in *IEEE Workshop Machine Learning for Signal Processing*, August 2007.

[33] T. Cormen, C. Leiserson, and R. L. Rivest, *Introduction to Algorithms.* The MIT Press, Cambridge, Massachusetts, 1990.

[34] S. Kullback, *Information theory and Statistics*. New York: Wiley, 1958.

[35] M. Padmanabhan and S. Dharanipragada, "Maximizing information content in feature extraction," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 512–519, July 2005.

[36] M. Bohanec and I. Bratko, "Trading accuracy for simplicity in decision trees," *Machine Learning*, vol. 15, pp. 223–250, 1994.

[37] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[38] J. Shen and G. Strang, "Asymptotic analysis of daubechies polynomials," *Proceedings of the American Mathematical Society*, vol. 124, no. 12, pp. 3819–3833, 1996.

[39] ——, "Asymptotics of daubechies filters, scaling functions, and wavelets," *Applied and Computational Harmonic Analysis*, vol. 5, pp. 312–331, 1998.

[40] A. M. Atto, D. Pastor, and A. Isar, "On the statistical decorrelation of the wavelet packet coefficients of a band-limited wide-sense stationary random process," *Signal Processing*, vol. 87, no. 10, pp. 2320 – 2335, 2007.

[41] A. M. Atto, D. Pastor, and G. Mercier, "Wavelet packets of fractional brownian motion: Asymptotic analysis and spectrum estimation," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 429–441, 2010.

[42] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[43] R. M. Gray, *Entropy and Information Theory*. Springer - Verlag, New York, 1990.

[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Interscience, New York, 1991.