Extracción de Parámetros de Señales de Voz usando Técnicas de Análisis en Tiempo-Frecuencia

F.A. Sepúlveda

Universidad Nacional de Colombia

Sede Manizales

Facultad de Ingeniería y Arquitectura

Departamento de Electricidad, Electrónica y Computación

Manizales, Colombia

2004

Extracción de Parámetros de Señales de Voz usando técnicas de Análisis en Tiempo-Frecuencia

F. A. Sepúlveda

Tesis para optar al título de
Magister en Automatización Industrial

Director
Prof. Germán Castellanos Domínguez

Universidad Nacional de Colombia
Sede Manizales
Facultad de Ingeniería y Arquitectura
Departamento de Electricidad, Electrónica y Computación
Manizales, Colombia

2004

Feature Extraction of Speech Signals using Time-Frecuency Analysis

by

F. A. Sepúlveda

A thesis submitted to the Posgraduate Program “Industrial Automation”
in partial fulfillment of the requirements for the Master Degree

Thesis Director
German Castellanos

Universidad Nacional de Colombia

Sede Manizales

Manizales, Colombia

2004

A mi hermosa familia,
por su amor inagotable.

Este trabajo se realiza en el marco del proyecto *Análisis acústico de voz y de posturas labiales en pacientes de 5 a 15 años con LPH corregido en la zona centro del país*, financiado por COLCIENCIAS.

Índice general

Índice de figuras

Índice de cuadros

Resumen

En este trabajo se presenta la extracción de características de señales de voz basada en transformada *wavelet*. Las características se pueden clasificar en los tipos acústico y de representación. Dentro de las características acústicas aparecen la frecuencia fundamental y la de medida de ruido de señales de voz. Para la estimación de la frecuencia fundamental se aplica un nuevo método, el cual usa la correlación de distancias entre las escalas de descomposición en lugar de usar la correlación de posiciones de máximos locales en las escalas. Para la obtener la medida de ruido de las señales de voz se usa un método basado en la transformada wavelet packet. Para la obtención de las características de representación se usan varia estrategias, la más simple de ella es usando la transformada *wavelet* diádica, y las otras se basan en el diccionario de bases generado a partir de la transformada *wavelet packet*, entre ellas *Local Discriminant Bases*.

Abstract

This work present methods for feature extraction of speech signals based on wavelet transform. The features can be organized in two categories, acoustic and representation. Present a new method for pitch estimation and use the wavelet packet transform for noise estimation. For extraction of representation features use the dyadic wavelet transform and schemes based on wavelet packet transform, por ejemplo, Local Discriminant Bases. This features are used for pathological voices classification and are evaluated using Linear Discriminant Analysis. As preprocessing technique we use an algorithm for voiced/unvoiced decision an later apply pitch estimation. The results are compared with other methods.

An improvement pitch detection algorithm based on the Wavelet Transform (WT) of speech signal is proposed. The method obtains a value of the fundamental frequency for each pitch period, is described and evaluated. In contrast with other methods, which chooses maximums if they occur in two adjacent wavelet coefficient scales, distances between adjacent local maximums are chosen for each scale. This method is computationally inexpensive and through real speech experiments shows that it is both accurate and robust to noise.

Capítulo 1

Introduction

El presente trabajo, recopila las actividades realizadas en torno a la Tesis de Maestría en Automatización Industrial realizada en el área de procesamiento de señales, específicamente en la extracción de características. Se desarrolla una metodología de extracción de características orientada a la clasificación de voces patológicas usando transformada *wavelet*.

La extracción de características consiste en determinar parámetros que describan el comportamiento dinámico de señales. Se trabaja con dos tipos de características: las características acústicas y las de representación. Dentro de las características acústicas se trabaja con la frecuencia fundamental y con la medida de ruido de señales de voz. Las características de representación no poseen sentido físico, pero recientemente se ha demostrado su potencial en tareas de clasificación de señales de voz.

La determinación de la frecuencia fundamental tiene sentido hacerla sobre aquellas porciones de señal del tipo sonoro, es decir donde participan las cuerdas vocales, por ello es necesario primero que todo realizar la segmentación sonora/sorda. Se presentan algoritmos para dicha tarea, y el mejor de ellos se compara con el presentado en el toolbox de Childers 2000. Sus resultados son comparables en desempeño pero el algoritmo basado en WT es de menor coste computacional.

En la literatura se menciona la *wavelet* madre usada para la estimación de la frecuencia fundamental pero muchas veces no se presenta una justificación clara del porque se usó determinada *wavelet*. Se plantea una metodología para la selección de la *wavelet* usando criterios teóricos y empíricos, desde el punto de vista de detección de singularidades.

Los coeficientes *wavelet* pueden ser usados a modo de características de señales, y reciente-

mente se ha mostrado su habilidad en tareas de clasificación. Se presenta una comparación de desempeño de características obtenidas a partir de varios esquemas en la clasificación de voces patológicas, obteniéndose buenos resultados para una cantidad pequeña de características.

En el *primer capítulo* se expone la fisiología del aparato fonador, aquí se brindan los conceptos teóricos a cerca del proceso de producción vocal por parte del aparato fonador humano, además, presenta las clasificaciones fisiológicas y la clasificación linguística de los fonemas.

En el *capítulo dos*, se presentan los conceptos de representación de señales en tiempo-frecuencia, inclinándose por la transformada *wavelet*, la cual resulta ser la más conveniente para la empresa que se aborda. En el *capítulo tres* se tratan tópicos de selección de bases *wavelet*, para las tareas de representación, clasificación y estimación del pitch.

En el *cuarto capítulo* se exponen algoritmos para estimación de voz, tanto acústicos como de representación. Se incluye la estimación de la frecuencia fundamental, la medida de ruido de señales de voz. Para la extracción de coeficientes *wavelet* se usa la transformada *wavelet* diádica, y la transformada *wavelet packet*.

Finalmente, en el capítulo cinco se describe el marco experimental realizado, se presentan las pruebas realizadas y sus correspondientes resultados. Se presentan tablas comparativas, que permiten analizar los resultados obtenidos para las diferentes pruebas realizadas.

Capítulo 2

Fisiología y lingüística del habla

2.1. Mecanismo de producción de la voz

La voz es producida por la excitación acústica de una cavidad variante en el tiempo, el tracto vocal, la cual es la región de la cavidad de la boca acotada por la cuerdas vocales y los labios. Los variados tipos de sonidos son producidos ajustando, tanto el tipo de exitación, como la forma del tracto vocal [77].

El mecanismo de producción de la voz puede modelarse por el sistema compuesto de tres etapas [27]:

Fuente: encargada de la generación de sonidos (pulmones, cuerdas vocales). Específicamente, el sonido puede pertenecer a uno u otro tipo: sonoro o insonoro. En donde se dice que un sonido es del tipo sonoro si participan las cuerdas vocales en su generación, de lo contrario se dice que es insonoro. La fuente de sonido del tipo sonoro puede ser modelado como un tren de pulsos o por ondas triangulares asimétricas las cuales son repetidas para cada periodo fundamental. De otra parte, el tipo de voz insonora puede ser modelado como un generador de ruido blanco [27].

Articulación (Modulador): le da forma y entonación a los sonidos que se están generando, comprende el tracto vocal el cual se puede modelar como una caja resonante que modifica (filtra) el sonido proveniente de las cuerdas vocales [54].

Radiación: corresponde a la parte final de las cavidades oral y nasal, por donde se expulsa el sonido.

Todo el aparato vocal humano está compuesto por los pulmones, la tráquea, la laringe, la faringe y las cavidades oral y nasal. El mecanismo de producción de la voz se inicia en los pulmones; el aire sale expulsado de ellos hacia la laringe (atravesando la traquea y la glotis) a diferente presión en función del sonido que se desea generar. La glotis separa las cuerdas vocales y se mantiene abierta mientras se respira, pero en el momento de producir sonidos se va estrechando de manera intermitente. La velocidad con la que las cuerdas vocales se abren y se cierran está ligada con lo que se conoce como *frecuencia fundamental*. Tras superar la glotis, el aire se acerca al tracto vocal, el cual varía su forma dependiendo de los sonidos a generar. El tracto vocal es una caja de resonancia , cuya forma, y por lo tanto su respuesta, varían de acuerdo a la posición de los órganos articuladores (lengua, labios, mandíbula, velo del paladar) [54]. Las resonancias producidas tienen su energía concentrada alrededor de determinadas frecuencias del espectro, a las que se refiere como *formantes* [19].

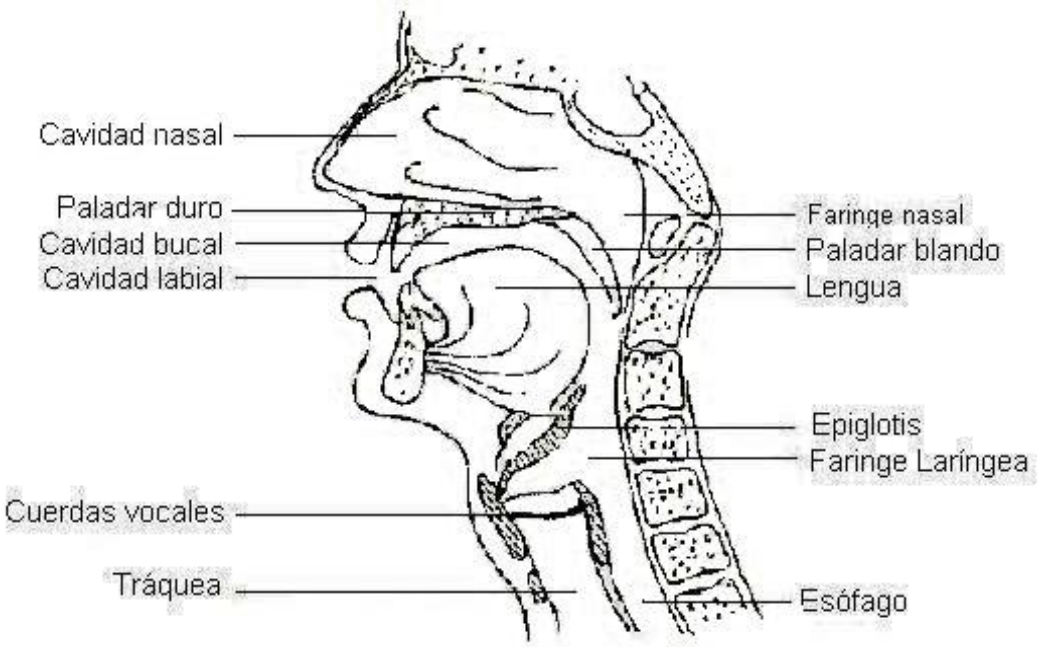

Figura 2.1: Aparato de fonador humano

Uno de los modelos más comunes de producción de la voz corresponde al modelo *fuente-filtro*, según el cual, las señales de voz son capturadas por micrófonos, los cuales responden

a cambios de presión en el aire. La presión del aire al salir de los labios $P_L(z)$, es obtenida de la forma:

$$P_L(z) = U_G(z)V(z)Z_L(z) \tag{2.1}$$

en donde $Z_L(z)$ corresponde a la impedancia presente en los labios, $V(z)$ es la función de transferencia del tracto vocal y $U_G(z)$ corresponde a la representación discreta en frecuencia de la señal de excitación proveniente de la faringe $u_G[n]$. En el caso de los sonidos del tipo sonoro la señal $u_G[n]$ dentro del modelo 2.1 corresponde a un tren de impulsos en convolución con el pulso glotal $g[n]$ y para voces del tipo sordo la señal $u_G[n]$ corresponde a ruido blanco gausiano .

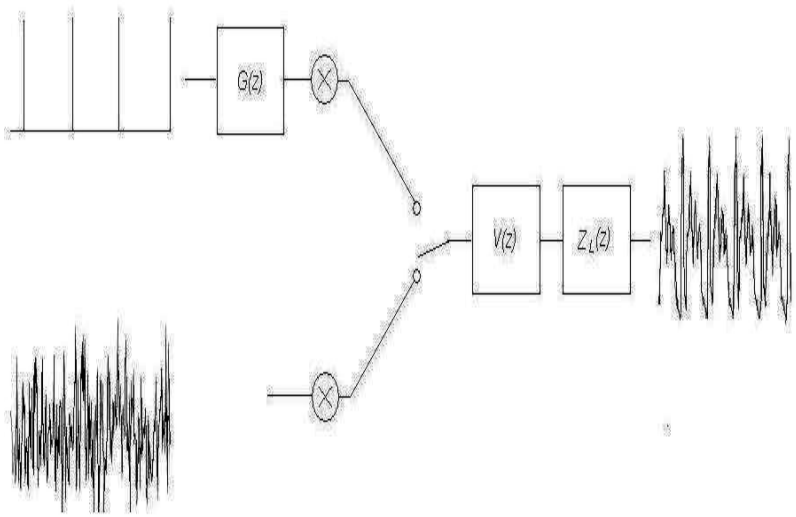

Figura 2.2: Modelo de producción de la voz.

2.2. El sistema auditivo humano

El oído humano está habilitado para desarrollar variadas tareas de procesamiento de las señales que percibe. Por ejemplo, posee mecanismos auditivos para realizar seguimiento de señales en ambientes ruidosos. Los cambios finos de intensidad y de frecuencia pueden ser medidos por el oído [29]. El oído procesa la señal de presión acústica transformándola

en patrones de vibraciones mecánicas en la membrana basilar, y representando dicho patrón por una serie de pulsos a ser transmitidos por el nervio auditivo [90].

El oído humano tal como se muestra en la figura (2.2), se divide en tres secciones: El oído externo, el oído medio y el interno. El oído externo consiste de la parte visible externa y el canal auditivo externo que forma un tubo a través del cual el sonido viaja. Este tubo tiene alrededor de 2.5 cm de largo y está cubierto por el tímpano cerca de su final. Cuando las variaciones de presión alcanzan el tímpano, este vibra, y transmite las vibraciones a huesos ubicados al otro lado del tímpano. El oído medio es una cavidad llena de aire de aproximadamente 6 cm^3 de volumen, en el cual se encuentran tres huesecillos, estrechamente interacoplados (yunque, estribo, martillo). Está ampliamente aceptado que la principal función del oído medio es la adaptación de impedancias entre los dos medios tan dispares que separa. Finalmente, se llega al oído interno, lugar donde se encuentra la cóclea, en forma de espiral y con una longitud aproximada de 35 mm. En el interior de la cóclea existe una membrana gelatinosa, llamada *membrana basilar*, que tiene un grosor mayor donde la cóclea es más estrecha. Conectados a ella hay decenas de millares de terminaciones nerviosas.

Dentro de la cóclea, una transformación de frecuencia a posición ocurre a lo largo de la membrana basilar, donde cada region de la membrana basilar es excitada por diferentes componentes de frecuencia de la señal de entrada. A medida que las ondas de sonido entran a la cóclea, una diferencia de presión es aplicada a través de la membrana basilar y el respectivo patrón de movimiento se desarrolla a lo largo de la membrana basilar en forma de ondas viajeras, las cuales se mueven desde la base hacia el ápice de la cóclea con amplitudes que se incrementan gradualmente hasta un cierto punto donde decrecen en forma abrupta. La posición donde una onda viajera alcanza su amplitud pico está fuertemente correlacionada con las propiedades físicas de la membrana basilar, donde las frecuencias altas tocan cerca a la base, la cual es angosta y rígida, y las señales de baja frecuencia tocan cerca al ápice donde la membrana basilar es ancha y suave. De esa manera, la membrana basilar puede tomarse como un analizador de Fourier básico o un analizador basado en un banco de filtros pasabanda que dividen la señal compleja en sus componentes de frecuencia [48].

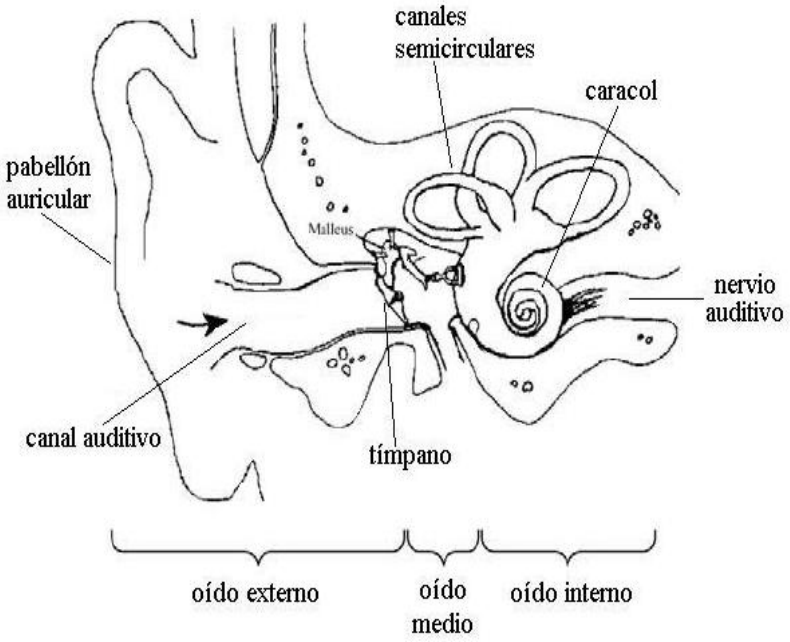

Figura 2.3: Aparato auditivo

2.3. Clasificación de los sonidos

Los sonidos se pueden clasificar según los siguientes criterios [28]:

1. La acción de las cuerdas vocales:

   a) *Sonoros*, si la cuerdas vocales se aproximan y comienzan a vibrar.

   b) *Sordos*, si las cuerdas vocales se acercan entre sí pero sin llegar a vibrar.

2. La acción del velo del paladar

   a) *Orales o Bucales*, si se encuentra adherido a la pared de la laringe. El aire sale sale solo entonces por la cavidad bucal. (/s/,/*b*/,/*p*/).

   b) *Nasales*, si parte del aire sale a través de la cavidad nasal.(/*n*/,/*m*/,/*g*/).

3. El modo de articulación, el cual puede definirse como la posición de los órganos articulatorios en referencia a su apertura o cierre, corresponde a la siguiente clasificación:

a) *Vocales*, si está en posición abierta. En general los sonidos vocálicos se caracterizan por su considerable amplitud.

b) *Consonantes*

- Oclusivas, cuando hay un cierre completo de los órganos articulatorios (/*p*/,/*t*/,/*k*/,/*b*/,/*d*/, /*g*/, /*m*/, /*n*/)
- Fricativas, cuando el sonido se forma a partir del estrechamiento de dos órganos articulatorios, en diferentes posiciones del tracto vocal sin que éstos nunca lleguen a juntarse. (/*f*/,/*s*/,/*x*/)
- Africadas, cuando tras un cierre completo de los órganos articulatorios sigue una pequeña apertura por donde se deslizará el aire contenido en un primer momento. (/*c*/,/*j*/).
- Nasales, cuando está abierta la cavidad nasal. (/*n*/,/*m*/,/*g*/).
- Líquidas, que pueden ser a su vez, Laterales (/*l*/), si el aire sale por un lado o ambos de la cavidad vocal,y Vibrantes (/*y*/,/*w*/,/*r*/) si se produce una vibración del ápice de la lengua.

4. El lugar de la articulación, corresponde a la siguiente división:

- *Bilabiales*, si se juntan los labios (/*p*/, /*b*/, /*m*/).
- *Labiodentales*, si se situan los dientes superiores en el labio inferior (/*f*/,/*v*/).
- *Interdentales*, si se coloca la lengua entre los dientes (/*t*/,/*d*/).
- *Alveolares*, si se pone la lengua adosada en las encías (/*s*/,/*z*/).
- *Palatales*, si se coloca la lengua adherida a la parte media y anterior del paladar duro (/s/,/z/).
- *Velares*, si se ubica la lengua contra el velo del paladar (/*k*/).
- *Glotales*, si se mantienen fijas y próximas las cuerdas vocales.

2.4. Patología de la voz

Las particularidades anatómicas y fisiológicas del tracto vocal en general, están determinadas por diferentes características cuyas variaciones definen la tipicidad (en el sentido de la normalidad del hablante referida a condiciones dadas) o atipicidad de la voz. Las patologías están definidas para cambios o variaciones fuera de los limites determinados como normales en la producción de voz. Se conoce una gran cantidad de alteraciones de la voz y el habla, que de manera significativa se reflejan en la naturaleza física de la señal.
Por cuanto, la presencia de patologías en los pliegues vocales puede causar cambios significativos en los patrones de vibración normales de los mismos, que a su vez desmejoran la calidad de la producción vocal, en la práctica, tiene importancia el análisis de las anomalías congénitas, o aquellas que hayan sido adquiridas, pero que de manera evidente influyan en la particularidad de la voz y el habla. Un ejemplo de esto puede ser las alteraciones de la voz debidas a la patología del aparato de voz periférico, así, cualquier cambio en la laringe condiciona fuertes perturbaciones en las funciones de la producción vocal, las cuales pueden ser clasificadas en dos grupos, aquellas que generan la disfonía y las que llevan a la afonía [83].

2.4.1. Disfonía

Se define como la alteración de una o varias de sus cualidades y características normales. En [26] se define la disfonía como un trastorno de la función vocal que frecuentemente se manifiesta por la alteración de uno o varios parámetros de la voz, como son: el timbre, la intensidad, y altura tonal [26]. Algunas de las causas de la disfonía son:

- *Laringitis aguda.* Ocurre por la inflamación de las cuerdas vocales debido a infección viral o al uso excesivo de la voz.
- *Nódulos de cuerdas vocales.* Es el engrosamiento del epitelio de la mucosa del repliegue vocal [47] que aparece en personas con mal uso vocal, que hablan muy alto, durante demasiado tiempo, o con mala técnica de emisión vocal.

- *Reflujo gastroesofágico.* El reflujo de material gástrico, sobre todo durante la noche, puede producir irritación de las cuerdas vocales y disfonía.
- *Parálisis de cuerdas vocales.* Por afectación del nervio recurrente debido a cirugía de la tiroides o compresión, consecuencia de tumoraciones, o sin causa aparente. Otras causas pueden ser también alergias o traumas de la laringe.

2.4.2. Afonía

Las razones de su aparición se deben a problemas en el sistema nervioso central, así como en patologías de la laringe. En este caso la voz no se genera, y el habla es posible sólo en forma de susurro. La afonía se debe a la parálisis y/o cortes en los músculos de la laringe, debido a la afección de la corteza cerebral o del cerebelo, además, en caso de problemas de infecciones y traumas del nervio inferior de la laringe o en alguna de sus ramificaciones. Como resultado de la parálisis de los músculos que sirven para la contracción y dilatación de la laringe, las cuerdas vocales no se cierran completamente y la voz desaparece. Se ha encontrado que en muchos pacientes, las disfunciones en la laringe están caracterizadas por [31]:

- Incremento en el grado de ronquera debido a que la voz de estos pacientes contiene componentes de ruido.
- Grandes variaciones en el pitch y las amplitudes pico del mismo.
- Quebrantes en la generación del pitch durante la emisión de vocales sostenidas.
- Presencia de componentes subarmónicos en el espectro de la vocal.
- Distorsión en la forma de los pulsos del pitch.
- Presencia de componentes de ruido de alta frecuencia.

En [64], se propone la clasificación de las voces disfuncionales como se detalla en la tabla:

Nombre	*Síntomas*
Ronquido	se refleja como una vibración aperiódica de las cuerdas vocales
Fatiga vocal	se presenta cansancio después de un habla prolongada
Soprosidad	hay incapacidad de pronunciar frases completas sin parar y sin precisar el re-abastecimiento de aire para continuar hablando
Extensión fonatoria reducida	síntoma generalmente asociado a cantores que presentan dificultades en producir notas que antes no tenían
Afonía	ausencia de voz
Quiebres de frecuencia	se presentan saltos periódicos de voz y de quiebres de voz. La voz parece fuera de control, o el paciente cuenta que nunca se sabe lo que pronunciará
Voz tensa/ comprimida	presentan dificultad al hablar. Esto puede incluir inhabilidad para hacer que una vocalización comience
Temblor	percepción de que su voz está sacudida o temblosa. Hay incapacidad de producir un sonido constante sustentado

Tabla 2.1: Clasificación de las voces disfuncionales

2.4.3. Labio y/o paladar hendido

El nacimiento de niños con algún problema de fisuras congénitas sirve de catalizador de un sinnúmero de problemas emocionales. Existen determinadas especialidades cuyo concurso se hace necesario para la evolución de dichos pacientes. Es necesario el intercambio de pareceres entre cirujanos especialistas en cirugía maxilo-facial y anormalidades otológicas, patólogos especializados en problemas del habla, ortodoncistas y prosodontistas, intercambio que deberá efectuarse de un modo periódico para perfilar el plan de tratamiento para cada caso en particular [92].

La fisura del labio y del paladar representan una de la deformidades congénitas más frecuentes, siendo la segunda en presentarse después del pie equinóvaro (pie zambo). Su frecuencia de aparición es mayor en los individuos de raza caucasiana (1,34 por cada 1000 nacimientos), que en individuos de raza negra (0.41 por cada 1000 nacimientos) [92].

Cuando hay paladar hendido, la voz que se produce es excesivamente nasal [13]. Esto es, se reduce la precisión de las consonantes oclusivas /*p*/, /*b*/, /*t*/, /*d*/, /*g*/) , fricativas /*s*/,/*z*/,/*f*/,/*v*/,/*d*/, /*sh*/) y africadas (/*ch*/) por el escape nasal. En el caso del labio hendido corregido, los sonidos que probablemente se afectan son los que requieren el cierre, arqueamiento y extensión de los labios /*p*/,/*b*/,/*m*/,/*u*/,/*i*/ [13].

El niño con paladar hendido tiene el riesgo de sufrir de problemas del habla y lenguaje relacionados con la afectación auditiva, déficit de sensaciones bucales, problemas sociales y emocionales y retardo en el desarrollo. No puede exagerarse la importancia de una intervención temprana [13].

2.4.4. Incompetencia velofaríngea

Se usa el término incompetencia velofaríngea(IVF) para denotar algún tipo de función velofaríngea anormal en la cual el velo del paladar y las paredes faríngeas laterales y posteriores fallan al separar la cavidad oral de la cavidad nasal. La IVF puede ser causada por deficiencias estructurales, afecciones neurológicas, e interferencias mecánicas en el cierre velofaríngeo.

El tracto vocal es una cavidad resonante, el acoplamiento y desacoplamiento de las cavi-

dades oral y nasal, dadas por la función velofaríngea, son necesarias para producir algunos tipos de sonidos y voz inteligible. El velo y la faringe actúan como válvulas que selectivamente encauzan el flujo de aire y la energía acústica, bajo diferentes presiones, a través de las cavidades oral y nasal. Para producir sonidos nasales, la válvula velofaríngea encauza el flujo de aire a través de la cavidad nasal. Similarmente, para los sonidos orales, la válvula cierra la cavidad nasal y selectivamente transmite el flujo de aire y energía a través de la cavidad oral. Además, para obtener voz normal, se debe tener una función velofaríngea rápida y competente, adicionada a la capacidad de producir flujos de aires apropiados y presiones adecuadas [35].

En la evaluación clínica los ojos y los oídos del profesional calificado son las primeras herramientas de diagnóstico. Como la medida más válida del adecuado funcionamiento del mecanismo velofaríngeo es el sonido producido durante la voz cotidiana, simplemente escuchándolo, se puede juzgar si se requieren futuras mediciones de la actividad velofaríngea. Cuando la voz de un paciente se desvía de la normal, una completa evaluación clínica es desarrollada. El exámen clínico procede con una examinación intraoral para medir la integridad estructural, el comportamiento reflexivo y el comportamiento fonético voluntario. Las características de producción de la voz son entonces valoradas, prestando especial a las siguientes características [35],

- Calidad de la resonancia. Es valorada mediante el chequeo de la presencia de vibración nasal durante la producción de vocales(/i/,/u/, y /e/).
- Flujos de aire. La presencia o ausencia de emisión nasal en los flujos de aire es la segunda característica de la producción de la voz que es valorada.
- Presión del aire. Se valora la adecuada presión en la cavidad oral.
- Existencia de mecanismos compensatorios. Pacientes con paladar hendido desarrollarán mecanismos compensatorios de articulación en diferentes áreas anatómicas que compensarán la ausencia o el pobre desempeño de la válvula velofaríngea.

La información invaluable obtenida a través del análisis clínico puede ser complementada por medio de mediciones objetivas obtenidas a través de la valoración instrumental. Las

técnicas instrumentales se dividen en los tipos directo e indirecto, por la forma de trabajar el equipo [19].

Al tipo de instrumentos indirectos pertenecen el Nasofaringoscopio, imágenes de resonancia magnética y el videofluoroscopio, entre otros. Las técnicas indirectas usadas para estudiar el mecanismo velofaríngeo son aquellas que proveen inferencia a través de datos tomados acerca de la estructura y cinemática del mecanismo velofaríngeo, por ejemplo la fototransducción. Otros tipos usan medidas indirectas tales como medidas aerodinámicas, acústicas, de presión del sonido y de espectrografía [35].

Los cambios cuantitativos en la resonancia nasal que ocurren con diferentes grados de desacoplamiento entre las cavidades oral y nasal proveen información acerca de la presencia o grado de la incompetencia velofaríngea. Para dichos pacientes un excesivo nivel de energía nasal es observada en la fonación de sonidos del tipo sonoro. En niños con obstrucción velofaríngea, se observa la falta de energía nasal en la fonación de sonidos del tipo nasal.

La hiper-nasalidad que es producida por el exceso de resonancia de la cavidad nasal, es generalmente mal clasificada dentro de los desórdenes de la voz. Realmente es un desorden de resonancia, los desórdenes de voz son causados por disfunciones de la laringe. La hipernasalidad es causada por una disfunción del mecanismo velo-faríngeo [13].

Todas las vocales son pronunciadas con el velo del paladar dirigido hacia a arriba, bloqueando la salida del aire hacia la cavidad nasal. Si el velo se baja durante la producción de las vocales, ocurre resonancia en la cavidad nasal, así como también en la cavidad oral, dándole un timbre particular a las vocales lo cual permite percibir cualidades nasales. Individuos con paladar hendido después de la cirugía o, después del acoplamiento de la alguna prótesis, requieren entrenamiento en el control del velo del paladar de tal forma que estén en capacidad de pronunciar sonidos explosivos, entre otros. Generalmente los individuos con IVF substituyen los sonidos explosivos por paros glotales [54].

Capítulo 3

Representación no-estacionaria de señales de voz

Una señal del tipo no-estacionaria es aquella cuyo contenido frecuencial varía respecto al tiempo. La transformada de Fourier ha sido ampliamente usada para el análisis de señales estacionarias, pero no puede usarse para el análisis frecuencial local respecto del tiempo.
Las señales de voz, en general, son del tipo no estacionario. Una representación adecuada de señales no estacionarias requiere del análisis frecuencial a nivel local respecto al tiempo, dando como resultado el análisis de señales en tiempo-frecuencia [25].
Dentro de la familia de las transformadas en tiempo-frecuencia, las que más se destacan son la transformada enventanada de Fourier, también conocida *Short Time Fourier Transform, STFT*, la transformada de *Wigner-Ville* y la transformada *wavelet*.

3.1. Representaciones en Tiempo-frecuencia

3.1.1. Transformada enventanada de Fourier

La transformada de Fourier, denotada por

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int e^{-i\omega t} f(t) dt \tag{3.1}$$

entrega el contenido frecuencial de la señal $f(t)$, pero no puede dar información frecuencial localizada en el tiempo, es más, los transitorios de alta frecuencia no podrán ser leídos fácilmente usando dicha transformada. La localización en el tiempo puede ser alcanzada

tomando secciones de la señal mediante el uso de ventanas, como se ve en la figura (3.1), para luego obtener la transformada de Fourier por cada ventana [24]; hay muchas formas de ventanas disponibles, por ejemplo Hanning, Hamming, coseno, Kaiser y Gaussiana, donde se considera a la STFT con ventana Gaussiana como la transformada Gabor [8]. Una ventana real y simétrica $g(t) = g(-t)$ es trasladada por u y modulada a la frecuencia ξ, para obtener el átomo de tiempo-frecuencia de la STFT:

$$g_{u,\xi}(t) = e^{i\xi t} g(t-u) \tag{3.2}$$

Una transformada lineal de tiempo-frecuencia correlaciona la señal con una familia de formas de onda que tienen su energía bien concentrada tanto en tiempo como en frecuencia, éstas formas de onda son los átomos de tiempo-frecuencia [55], por lo que la transformada enventanada de Fourier de $f \in \mathbf{L}^2(\mathbb{R})$ está dada por,

$$Sf(u,\xi) = \langle f, g_{u,\xi} \rangle = \int_{-\infty}^{\infty} f(t) g(t-u) e^{-i\xi t} dt \tag{3.3}$$

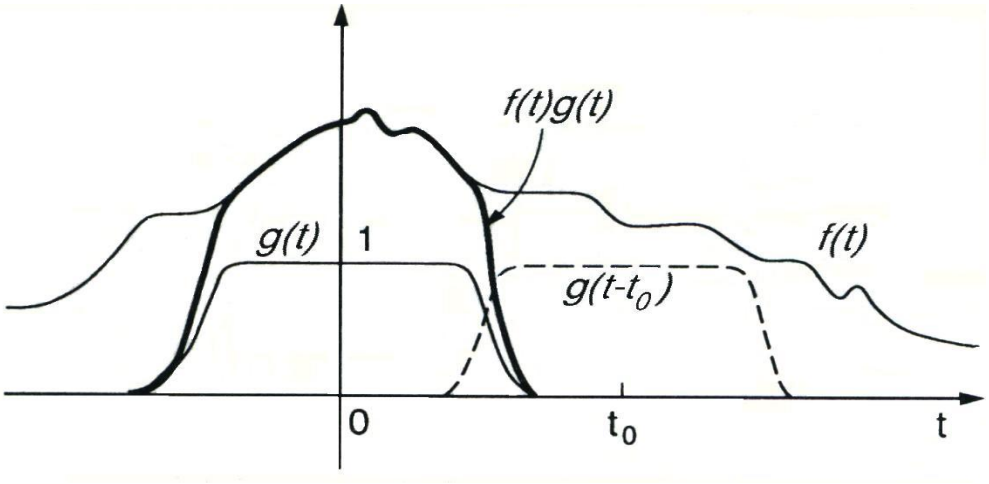

Figura 3.1: Dos etapas del árbol de análisis de dos bandas

Se puede obtener una distribución de la energía de la señal f en el plano tiempo-frecuencia

usando el espectrograma, denotada por [24],

$$P_s f(u,\xi) = \mid Sf(u,\xi) \mid^2 = \int_{-\infty}^{\infty} \mid f(t)g(t-u)e^{-i\xi t}dt \mid^2 \tag{3.4}$$

la expresión (3.4) mide la energía de la señal f en la vecindad de (u,ξ) [55].

3.1.2. Transformada de Wigner-Villle

Motivado por la medición de frecuencias instantáneas, en 1948 Ville introdujo para el procesamiento de señales una forma cuadrática que mide la energía local en los dominios conjuntos de tiempo y frecuencia, definida por:

$$W_f(w,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \bar{f}(t-\tau/2)f(t+\tau/2)e^{-iw\tau}d\tau \tag{3.5}$$

donde $W_f(w,t)$ es una función de t y w. El complejo conjugado $\bar{f}(t)$ es introducido para generalizar el análisis de señales complejas [60]. La transformada o distribución de Wigner-Ville es una correlación cruzada de la señal consigo misma para un corrimiento de tiempo y de frecuencia dados. Si la energía de f está bien concentrada alrededor de t_0 y en la frecuencia alrededor de ω_0, entonces $W_{ff}(\omega,t)$ tendrá la energía concentrada en (t_0,ω_0) [55]. La transformada de Wigner-Ville carece de una propiedad fundamental de una densidad de energía: la positividad, lo cual conlleva a problemas de interpretación [60].

La distribución de Wigner-Ville es no-lineal. Esto significa que la distribución de la suma de dos señales no es simplemente la suma de las distribuciones Wigner-Ville de cada una de las señales [25]; en lugar de ello la distribución de Wigner-Ville de la suma de dos funciones está dada por

$$W_{f+g}(t,\omega) = W_f(t,\omega)W_g(t,\omega) + 2ReW_f(t,\omega)W_g(t,\omega) \tag{3.6}$$

3.1.3. Transformada wavelet continua

Una *wavelet* corresponde a una función que tiene su energía concentrada en tiempo [84]. Las *wavelets* se emplean en el análisis de fenómenos no estacionarios o variantes en el

tiempo [81], entre otros. Una *wavelet* madre es una forma de onda, donde sus versiones trasladadas y escaladas vienen a formar las funciones base en la representación *wavelet*. La transformada *wavelet* descompone señales usando *wavelets* dilatadas y trasladadas, en donde una *wavelet* es una función que cumple las siguientes propiedades:

– media cero

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \tag{3.7}$$

Lo cual implica que la *wavelet* debe tener al menos una oscilación [4].

– Energía finita

$$\int_{-\infty}^{\infty} \mid \psi(t) \mid^2 dt \quad < \infty, \quad \psi \in \mathbb{L}^2(\mathbb{R}) \tag{3.8}$$

– C_ψ, denominada *Condición de Admisibilidad*, depende solo de la función *wavelet* madre $\psi(t)$, y se debe cumplir que

$$C_\psi = 2\pi \int_{-\infty}^{\infty} |\hat{\psi}(\xi)|^2 |\xi|^{-1} d\xi < \infty \tag{3.9}$$

donde $\hat{\psi}$ corresponde a la transformada de Fourier de ψ [20] [17].

Los coeficientes de la transformada *wavelet* (WT) se obtienen a partir de la correlación de la señal con los átomos de tiempo de frecuencia de la WT, donde cada átomo de tiempo-frecuencia $\psi_{u,s}$ corresponde a la versión trasladada y escalada de la *wavelet* madre ψ,

$$\psi_{s,u}(t) = \frac{1}{\sqrt{s}} \psi(\frac{t-u}{s}) \tag{3.10}$$

donde s es el parámetro de escala y u es el parámetro de traslación. La *Transformada Wavelet Continua (Continuos Wavelet Transform - CWT)* de una función $f(t)$ se define como:

$$Wf(u,s) = \langle f(t), \psi_{u,s} \rangle = \int_{-\infty}^{\infty} f(t) \frac{1}{\sqrt{s}} \psi(\frac{t-u}{s}) dt \tag{3.11}$$

donde $\langle , \rangle$ corresponde al producto interno en el espacio $\mathbf{L}^2(\mathbb{R})$.

La señal original puede ser reconstruida a partir de su CWT utilizando la expresión de reconstrucción [25]

$$f(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \langle f(\tau), \psi_{s,u}(\tau) \rangle \psi_{s,u}(t) \frac{dsdu}{s^2} \tag{3.12}$$

donde la constante C_ψ es la constante de admisibilidad, dada por (3.9).

La diferencia entre la transformada *wavelet* y la STFT radica en la forma de las funciones de análisis $g_{u,\xi}(t)$ y $\psi_{u,s}$. Las funciones $g_{u,\xi}$ en la STFT consisten todas de la misma función envolvente g, trasladada adecuadamente, pero existiendo dentro de sí oscilaciones de alta frecuencia. Todas las $g_{u,\xi}$ independiente del valor de ξ, poseen la misma apertura de ventana. En contraste las *wavelets* $\psi_{u,s}$ poseen aperturas de ventana que se adaptan a su frecuencia, tal como se ilustra en la figura (3.2); las funciones $\psi_{u,s}$ de muy alta frecuencia son muy angostas mientras que $\psi_{u,s}$ de baja frecuencia son más amplias. Como resultado de dicha propiedad la WT está en mejores condiciones para detectar transitorios de alta frecuencia, tales como singularidades [24].

Ambos métodos STFT y Wigner-Ville son de banda constante. Sus algoritmos requieren una transformación del dominio del tiempo al dominio de la frecuencia usando la transformada discreta de Fourier (Discrete Fourier Transform-DFT) [93], lo cual genera coeficientes de Fourier para frecuencias a igual separación. El ancho de banda para cada término frecuencial es el mismo, en contraste, la WT permite que el ancho de banda se pueda ajustar [60].

Una variedad de consideraciones prácticas se pueden realizar al algoritmo de la STFT para mejorar la definición del mapa de tiempo-frecuencia que genera. Concretamente, en el incremento de la definición en frecuencia, los datos de la ventana de análisis pueden ser aumentados adicionando ceros, (*zeropadding*), con lo cual se incrementa el número de armónicos generados por la FFT. Para mejorar la resolución en el tiempo, las ventanas sucesivas de análisis se sobrelapan, de tal manera que la posición del centro de la ventana se incrementa a paso pequeños. Dichas modificaciones generan más puntos para dibujar en el mapa de tiempo-frecuencia, mejorando su apariencia y ayudando a su interpretación.

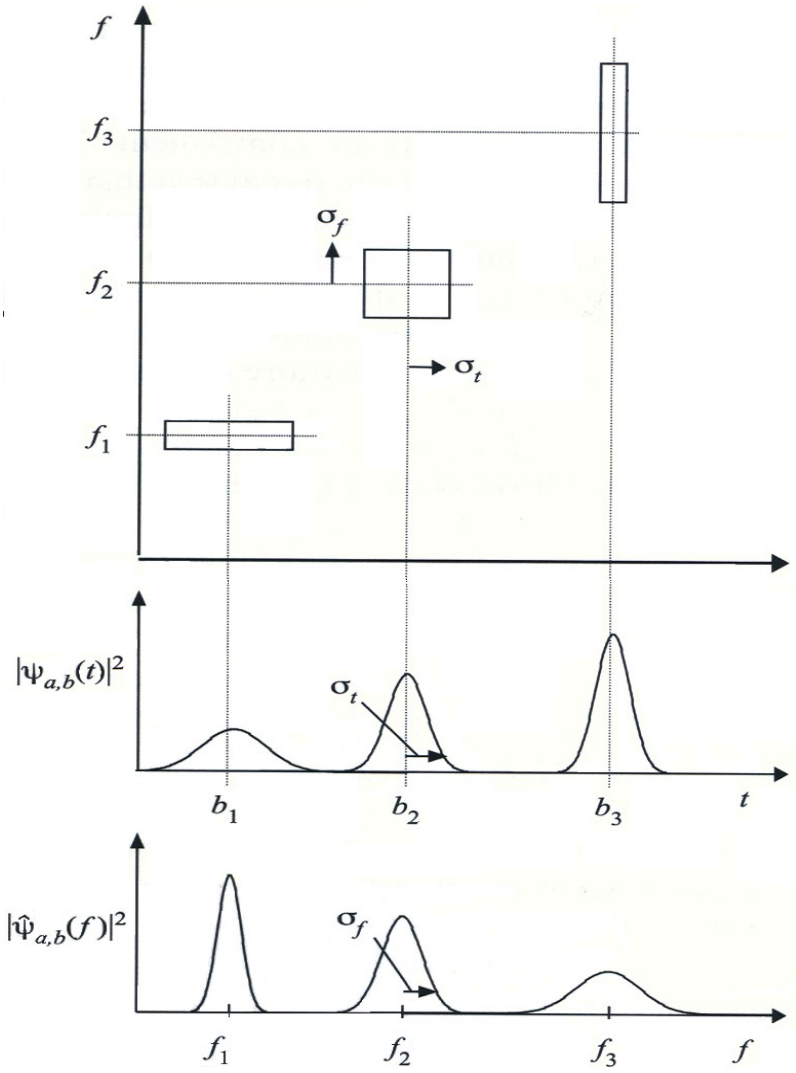

Figura 3.2: Distribución de los átomos de tiempo-frecuencia de la transformada *wavelet*

Sin embargo, la resolución en frecuencia no es modificada debido a que esta depende del tamaño de la ventana de análisis y del número de muestras de la ventana [60].
Existen variados tipos de transformada *wavelet* (WT), pero todas ellas empiezan en la fórmula básica (3.11). Las dos grandes categorías corresponden a la transformada *wavelet* Continua (CWT, *Continuos Wavelet Transform*) y la transformada *wavelet* discreta (DWT, *Discrete Wavelet Transform*). Dentro de las transformadas *wavelet* discretas se distinguen los sistemas discretos redundantes y las transformadas *wavelet* de bases ortonormales [24].

3.2. Transformada wavelet discreta

El cálculo de los coeficientes *wavelet* para cada escala posible es una tarea de alto coste computacional, y genera una enorme cantidad de datos, muchas veces innecesarios. Por lo tanto se plantea el cálculo de dichos coeficientes para determinados valores de escala y de posición.
Una señal $f(t)$ puede ser expresada como la descomposición lineal de la forma

$$f(t) = \sum_{l} a_l \phi_l(t) \tag{3.13}$$

donde a_l son los coeficientes de la expansión, y ϕ_l son una familia de funciones evaluadas en t.
En la formación del sistema *wavelet* se requieren dos parámetros, uno asociado a la traslación y otro asociado a la escala, de tal forma que (3.13) se convierte en [17]:

$$f(t) = \sum_{n} \sum_{j} a_{j,n} \phi_{j,n}(t) \tag{3.14}$$

En el análisis de señales complejas con estructuras de conformación que son muy variadas, tales como las señales de voz, es conveniente usar átomos de representación $\phi_{j,n}(t)$ cuya forma sea igualmente variada.

En la transformada *wavelet* continua, se considera la familia [55] [24]

$$\psi_{s,u}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t-u}{s}) \tag{3.15}$$

donde $u \in \mathbb{R}$, $s \in \mathbb{R}_+$ con $s \neq 0$, y ψ cumple la condición de admisibilidad dada por (3.9). Lo que se desea es restringir los parámetros s, u a valores discretos únicamente. Para s se busca un $s = s_0^j$, donde $j \in \mathbb{Z}$, y el paso de dilatación $s_0 \neq 1$ es fijo. La discretización de u se realiza tomando únicamente los enteros múltiplos de un valor fijo u_0 (arbitrariamente se escoge $u_0 > 0$), donde u_0 es buscado apropiadamente, tal que $\psi(t - nu_0)$, lo que implica que $j = 0$, cubra por completo el eje de las abscisas.
Debido a que valores diferentes de j corresponden a *wavelets* de diferente amplitud, la discretización del parámetro de traslación u debería depender de j. Así, *wavelets* angostas (de alta frecuencia ó de escalas bajas) son trasladadas por pasos pequeños en orden a cubrir todo el rango de tiempo, mientras que *wavelets* amplias (de baja frecuencia ó de escalas altas) deberían ser trasladadas por pasos grandes. u estaría dado por $u = nu_0 s_0^j$, donde u_0 es un valor fijo, y $n \in \mathbb{Z}$ [24]. Las *wavelets* discretizadas quedarían de la forma,

$$\psi_{j,n}(t) = s_0^{-j/2}\psi(s_0^{-j}t - nu_0) \tag{3.16}$$

Al plantearse el problema de la discretización de la transformada surgen las preguntas:

- ¿Es posible caracterizar completamente una señal a partir de su transformada discreta?
- ¿Es necesario conocer todos los valores de la descomposición *wavelet* continua para reconstruir exactamente la señal que se analiza?

Si en la representación *wavelet* se tienen más coeficientes de los necesarios para una reconstrucción exacta, entonces se dice que dicha representación es redundante. El hecho de discretizar la transformada no implica que se elimine la redundancia en una representación, es decir, no necesariamente se cae en una representación ortogonal; y en variadas aplicaciones la propiedad de redundancia podría ser útil, en particular, en el presente trabajo se

usa una transformada *wavelet* discreta redundante para la obtención de la frecuencia fundamental de las señales de voz.

3.2.1. Marcos

Un marco está formado por una familia de vectores $\{\phi_m\}_{m\in\Gamma}$ que representan una señal f a partir de sus productos internos $\{\langle f, \phi_m\rangle\}_{m\in\Gamma}$ [55], donde Γ es un conjunto de índices que puede ser de tamaño finito o infinito. La familia de *wavelets* $\psi_{j,n}(t), n \in \mathbb{Z}, j \in \mathbb{Z}$ en $\mathbf{L}^2(\mathbb{R})$ constituye un marco si existen $A > 0$ y $B < \infty$ tales que para toda función $f \in \mathbf{L}^2(\mathbb{R})$ se cumple que:

$$A \leq \frac{\sum_{j,n} \mid \langle f(t), \psi_{j,n}(t)\rangle \mid^2}{\parallel f(t) \parallel^2} \leq B \tag{3.17}$$

Se observa que la energía normalizada de los coeficientes está acotada por los valores de A y B, es decir, ellos "enmarcan"la energía normalizada, dichos valores dependen de s_0 y u_0 y de la *wavelet* madre [8]. El marco será una base ortonormal si se cumple que $A = B = 1$ [17].

Teniendo en cuenta elecciones adecuadas de la función *wavelet* madre $\psi(t)$ y de los parámetros s_0, u_0; $\psi_{j,n}$ constituiría una base ortonormal para $\mathbf{L}^2(\mathbb{R})$. En particular, si se escoge $s_0 = 2, u_0 = 1$, entonces existe una función ψ, con buenas propiedades de localización de tiempo-frecuencia, de tal forma que

$$\psi_{j,n}(t) = 2^{-j/2}\psi(2^{-j}t - n) \tag{3.18}$$

constituye una base ortonormal para $\mathbf{L}^2(\mathbb{R})$.

Existen varias ventajas para que se requiera que las *wavelets* sean ortogonales. Las funciones de base ortogonal permiten el cálculo rápido de los coeficientes de expansión, y cumplen el teorema de Parseval, que permite la partición de la señal en el dominio *wavelet* [17]. Pero existen casos en donde las bases para el sistema *wavelet* no pueden, ser ortogonales. Para tales casos se suele usar un conjunto de *bases duales* $\tilde{\phi}_k(t)$, los cuales cumplen la relación,

$$\langle \phi_l(t), \tilde{\phi}_k(t)\rangle = \delta(l - k) \tag{3.19}$$

Debido a que este tipo de ortogonalidad requiere dos conjuntos de vectores, conjunto de expansión y conjunto dual, el sistema es llamado *biortogonal*. Para dicho sistema la función f quedaría representada de la forma,

$$f(t) = \sum_k \langle f(t), \tilde{\phi}(t) \rangle \phi_k(t) \tag{3.20}$$

Los filtros *wavelet* asociados a las bases ortogonales generalmente son de fase no lineal, pero al debilitar dicha condición en la construcción de los filtros, se consiguen filtro de fase lineal tanto para el análisis como para la reconstrucción [88]. A pesar de que un sistema biortogonal es más complicado y requiere almacenar un conjunto adicional al de expansión, el conjunto dual, dicho sistema es más general y las *wavelets* biortogonales poseen la ventaja de que sus filtros *wavelet* son de corta longitud y fase lineal, lo cual difícilmente podría conseguirse sin relajar la condición de ortogonalidad [53].

3.2.2. Análisis de multiresolución y banco de filtros

La resolución de una señal es una descripción cualitativa asociada con su contenido frecuencial. La versión sub-muestreada o la salida de un filtro pasa-bajas es usualmente una buena aproximación para muchas señales. La multiresolución es especialmente evidente en procesamiento de imágenes y visión artificial, donde la versión gruesa de una imagen es generalmente usada como una primera aproximación. El principio básico del análisis de multi-resolución consiste en representar una función o señal f como un límite de aproximaciones sucesivas, cada una de las cuales es una versión más fina de la función f. Estas aproximaciones sucesivas corresponden a diferentes niveles de resolución . Los sistemas de multi-resolución basados en bancos de filtros lo que hacen es tomar la entrada del sistema y dividirla en sub-secuencias usando bancos de filtros [32]. Esta técnica tiene trayectoria histórica en el área de voz [3], [67] y de hecho, los esquemas más convenientes de codificación de imágenes están basados en expansiones de bancos de filtros [5] [75] [30].

Definición 3.1. *(El Análisis de Multiresolución (MRA, Multi-resolution Analysis) consiste de una secuencia* $\{V_j : j \in \mathbb{Z}\}$ *de subespacios cerrados embebidos de* $\mathbf{L}^2(\mathbb{R})$ *que satisfacen las siguientes condiciones:*

1. $\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \cdots V_j \subset V_{j+1}$
2. $\bigcup_{m=-\infty}^{\infty} V_m$ *es denso en* $\mathbf{L}^2(\mathbb{R})$
3. $\bigcap_{m=-\infty}^{\infty} V_j = \{0\}$
4. $f(t) \in V_j$ *si y solamente si* $f(2t) \in V_{j+1}$ *para todo* $j \in \mathbb{Z}$
5. *Existe una función* $\varphi \in V_0$ *tal que* $\{\varphi_{0,n} = \varphi(t-n), n \in \mathbb{Z}\}$*, es una base ortonormal para* V_0 .

A la función φ se le denomina *función de escalamiento* o *wavelet padre*. Si $\{V_j\}$ es una multiresolución de $L^2(\mathbb{R})$ y si V_0 es el subespacio cerrado generado por las traslaciones enteras de la función φ, entonces se dice que φ genera el espacio de multiresolución. Algunas veces la condición 5 es debilitada asumiendo que $\{\varphi(t-n), n \in \mathbb{Z}\}$ es una base del tipo *Riesz* para V_0 [25]. Donde una *base Riesz* se define como el *marco* donde sus elementos son linealmente independientes [55].

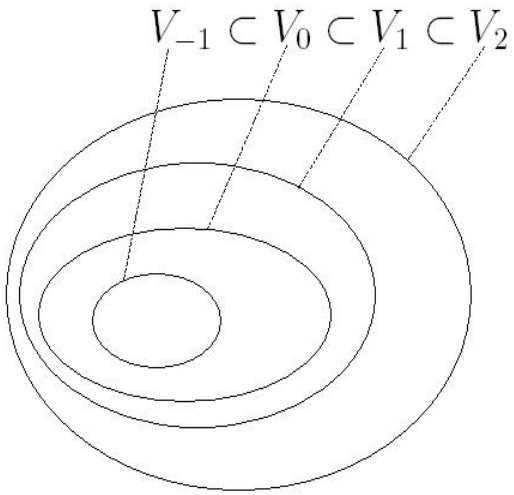

Figura 3.3: Trasformada Wavelet

Dado que $V_j \subset V_{j+1}$, se define W_j como el complemento ortogonal de V_j en V_{j+1} para cada $m \in \mathbb{Z}$ de tal forma que se tiene que [25],

$$\begin{aligned} V_{j+1} &= V_j \oplus W_j \\ &= (V_{j-1} \oplus W_{j-1}) \oplus W_j \\ &= V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_j \end{aligned} \tag{3.21}$$

Lo espacios W_j son denominados los espacios *wavelet* o espacios de detalle, y a los espacios V_j se les denomina espacios de aproximación.
Sea $P_j f$ la proyección del espacio V_j sobre la función f y sea $Q_j f$ la proyección de W_j sobre la función. La diferencia entre las dos aproximaciones sucesivas $P_j f$ y $P_{j+1} f$ está dada por la proyección ortogonal $Q_j f$ tal que [24]

$$Q_j f = P_{j+1} f - P_j f \tag{3.22}$$

Además los espacios W_j y W_k son mutuamente ortogonales para $j \neq k$ [17]. Para cada subespacio W_j existe una función ψ tal que

$$\psi_{j,n}(t) = 2^{j/2} \psi(2^j t - n) \tag{3.23}$$

constituye una base ortonormal para W_j, donde $n \in \mathbb{Z}$.

3.2.3. Transformada wavelet diádica

Para propósitos de estimación del *pitch* es suficiente tomar sólo unas cuantas escalas de la transformada *wavelet* continua, sin embargo, en el cálculo computacional, es preferible el empleo de algoritmos rápidos. En el presente trabajo se usó el algoritmo *Algoritme à Trous* o transformada *wavelet* diádica (DyWT, *Dyadic Wavelet Transform*) [55], el cual está basado en bancos de filtros. El algoritmo se ilustra en la figura 3.2.3.
En el desarrollo de la DyWT, la escala es muestreada a lo largo de secuencias diádicas $\{2^j\}, j \in \mathbb{Z}$, para hacer más rápidos los cálculos numéricos. La DyWT de $f \in \mathbb{L}^2$ está definida por:

$$W f(u, 2^j) = \int_{-\infty}^{\infty} f(t) \frac{1}{\sqrt{2^j}} \psi \left(\frac{t-u}{2^j} \right) dt = f \star \bar{\psi}_{2^j}(u) \tag{3.24}$$

siendo

$$\bar{\psi}_{2^j}(t) = \psi_{2^j}(-t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{-t}{2^j}\right)$$

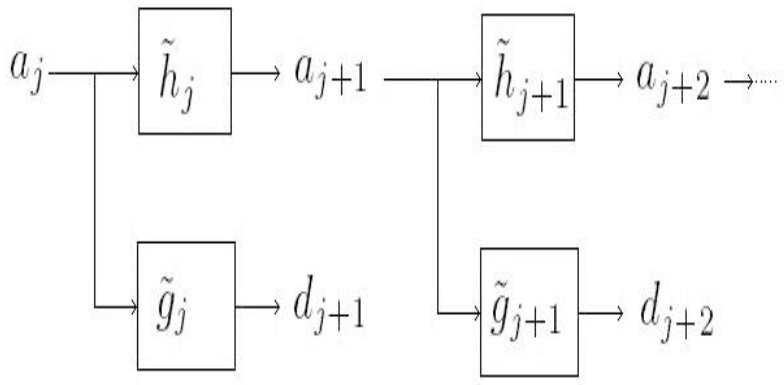

Figura 3.4: DWT: *Algoritme à Trous*

Históricamente, la CWT se creó primero. El enlace entre los bancos de filtros y el análisis de multiresolución sólo llegó a ser claro a finales de los 80's con el trabajo de S. Mallat [56]. Un ejemplo típico de un marco *wavelet* se consigue a partir de la discretización del parámetro de la escala s:

$$s_j = 2^{-j}, j = L, ..., J-1 \tag{3.25}$$

El marco consiste de dilataciones diádicas y traslaciones de la función madre,

$$\psi_{j,n}(t) = \psi(2^j t - 2^{j-J} n) \tag{3.26}$$

Si la *wavelet* madre se ajusta a un MRA, los coeficientes del marco se obtienen a partir de un algoritmo de banco de filtros a múltiples escalas, muy similar al algoritmo de la transformada rápida *wavelet* (FWT, *Fast Wavelet Transform*) basado en banco de filtros. Este algoritmo se consigue al omitir la etapa de submuestreo en el algoritmo clásico de la transformada *wavelet*. A la transformada presentada también se le conoce con los nombres de *Non-Decimated Wavelet Transform*, *Redundant Wavelet Transform* y *Stationary Wavelet Transform*.

En reconocimiento de patrones, es importante construir representaciones que sean invariantes a la traslación [55], [24]. Cuando un patrón es trasladado, sus descriptores numéricos podrían ser trasladados, pero no modificados. En tales casos una representación que dependa de la localización trae dificultades. La transformada *wavelet* continua y la transformada de Fourier proveen representaciones invariantes a las traslaciones [55]. La transformada enventanada de Fourier y la transformada *wavelet* discreta generan representaciones que son no invariantes a la traslación. La transformada *wavelet* diádica mantiene la invarianza a la traslación muestreando únicamente el parámetro escala de la CWT.
En muchos métodos existentes, los valores de energía o los coeficientes *wavelet* en sí mismos para todas las escalas o para las escogidas son tomados como características esenciales de la señal. Esto ayuda a sobrellevar el problema de que los coeficientes *wavelet* son no invariantes a corrimientos en el proceso de la señal [65]. En [12] se presenta una comparación de dichas transformadas para reconocimiento de fonemas.

3.2.4. Características de la wavelet

En técnicas de análisis de tiempo-frecuencia la representación depende del tipo de kernel usado. En variados trabajos [4, 14, 10] se trata de explotar tal propiedad para para obtener un máximo de rendimiento. Es por ello que se incluye a continuación las propiedades más importantes encontradas en la *wavelet* madre o su filtro asociado.

Momentos de desvanecimiento

Para medir la regularidad local de una señal, el uso de *wavelets* con soporte frecuencial angosto no es muy importante, sin embargo los momentos de desvanecimiento son cruciales [55]. Si la *wavelet* tiene n momentos de desvanecimiento, se puede demostrar que la transformada *wavelet* puede ser interpretada como un operador diferencial multi-escala de orden n [55].
Una de las propiedad de la WT es su capacidad para ser ciega a ciertos comportamientos regulares. Se sabe que una condición necesaria para que la *wavelet* exista es que su integral es zero, ecuación (3.7). Por lo tanto, dos funciones que difieren únicamente por una con-

stante tienen los mismos coeficientes *wavelet* de refinamiento: la WT no ve las constantes. De ser necesario, se puede hacer que sean ciegas a determinados comportamientos polinomiales de alto orden, imponiendo condiciones a los momentos de desvanecimiento [82]. Esto se ilustra de mejor manera mediante el concepto de la regularidad de Lipschitz.

La caracterización de la singularidad local de una señal o función se realiza mediante el exponente de Lipschitz, así si f tiene una singularidad en v, entonces el exponente de Lipschitz en v caracteriza dicho comportamiento singular. Una función es Lipschitziana con exponente α en la vecindad de v, si existe $K > 0$, y una expresión polonomial p_v de grado $m = \llcorner \alpha \lrcorner$ (el entero m más grande tal que $m \leq \alpha$) tal que

$$\vee t \in \mathbb{R}, \mid f(t) - p_v(t) \mid \leq K \mid t - v \mid^{\alpha} \tag{3.27}$$

Donde p_v es la expresión polinomial de Taylor en la vecindad v, la cual está dada por

$$p_v(t) = \sum_{k=0}^{m-1} \frac{f^{(k)}(v)}{k!}(t-v)^k \tag{3.28}$$

Ayudados por la expresión (3.28) se puede aproximar f a un polinomio p_v en la vecindad de v de la forma [55]:

$$f(t) = p_v(t) + \epsilon_v(t) \tag{3.29}$$

La transformada *wavelet* podría ignorar el polinomio p_v. Para tal propósito se usa una *wavelet* que tenga K *momentos de desvanecimiento*, definidos como:

$$\int_{-\infty}^{\infty} t^k \psi(t) \partial(t) = 0 \quad para \quad 0 \leq k < K \tag{3.30}$$

Se asume que la señal a analizar posee singularidades, y que entre singularidades la señal tiene un comportamiento suave. En éstos intervalos de comportamiento suave la señal puede ser aproximada adecuadamente de forma local mediante una expresión polinomial del tipo mostrado por (3.28), la cual está formada formada por la suma de monomios x^k ponderados. Dado que la *wavelet* posee K momentos de desvanecimiento, según (3.30), entonces para aquellos monomios de grado menor o igual a K se cumple que,

$$\langle x^k, \tilde{\psi}_{j,n} \rangle = 0, \quad k = 0, ..., K \tag{3.31}$$

por lo que se puede asegurar que los primeros K términos en la aproximación de Taylor de una función analítica no contribuyen a los coeficientes *wavelet* [42]. A partir de (3.29) y (3.31) se puede demostrar que,

$$Wf(j,n) = W\epsilon_v(j,n) \tag{3.32}$$

lo cual explica el porqué la transformada *wavelet* entrega máximos locales en aquellos instantes donde exista alguna singularidad.

Regularidad

La regularidad es una forma de medir la suavidad de una función, un alto grado de regularidad de una función en un punto indica que dicha función es suave en dicho punto. Una primera forma de medir la regularidad de una función en un punto consiste en verificar la existencia de sus derivadas (primera derivada, segunda derivada, ...). La regularidad de ψ tiene influencia en en el error introducido por umbralización de los coeficientes *wavelet*.
Cuando se reconstruye una señal a partir de sus coeficientes *wavelet*

$$f = \sum_{j=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \langle f, \psi_{j,n} \rangle \psi_{j,n},$$

el error ϵ adicionado al coeficiente $\langle f, \psi_{j,n} \rangle$ adicionará el componente *wavelet* $\epsilon\psi_{j,n}$ a la señal reconstruida. Si ψ es suave, entonces $\epsilon\psi_{j,n}$ será un error que varía suavemente a través de la señal [55]. Para aplicaciones de codificación de imágenes, un error suave es generalmente menos visible que un error irregular, inclusive si ellos tienen la misma energía. Para las *wavelets* de la familia *Daubechies* la regularidad está relacionada con el orden de la familia,p.e. $db1$, $db2$,...,dbn; un alto orden indica que la *wavelet* posee más regularidad. Las imágenes de mejor cualidad, desde el punto de vista de compresión, generalmente son obtenidas usando *wavelets* con carácter de continuidad diferencial que usando *wavelets* discontinuas, por ejemplo la *wavelet* Haar.
Los coeficientes *wavelet* (para el caso de la transformada *wavelet* discreta) son obtenidos por la convolución de la señal con los filtros, los cuales son obtenidos a partir de las

wavelets. Si la *wavelet* no poseen la suficiente regularidad, y por lo tanto los filtros tampoco, entonces la DWT podría producir coeficientes de amplitud alta para cambios pequeños de la señal, tal es el caso de la *wavelet* Haar.

Soporte compacto

El soporte de una función $f : \mathbb{R} \rightarrow \mathbb{C}$ corresponde a $\{x : f(x) \neq 0\}$, y es denotado por $supp(f)$. Una función posee *soporte acotado* si existen dos números reales a, b tal que $supp(f) \subset (a, b)$. Si el soporte acotado es cerrado se denomina *soporte compacto*. La siguiente expresión permite calcular el soporte de la *wavelet* y su filtro asociado a partir del filtro h [55].

Proposición 3.2. *La función de escalamiento φ tiene soporte compacto si y solo si h tiene soporte compacto y sus soportes son iguales. Si el soporte de h y φ es $[N_1, N_2]$ entonces el soporte de ψ es $[\frac{N_1-N_2+1}{2}, \frac{N_2-N_1+1}{2}]$*

Fase lineal

En forma más general, un filtro h es llamado de fase lineal si se cumple que [89],

$$\hat{h} = c' e^{jc\omega} h_R \tag{3.33}$$

donde c' es una constante compleja, c es un número real y h_R es una función real de w. Esta definición demuestra que la respuesta de fase del filtro puede ser lineal a tramos.
La respuesta de fase lineal de los filtros evita la distorsión producida por la fase no-lineal y mantiene la forma de la señal, lo cual es usualmente importante en aplicaciones de audio [89], imágenes [89], [66], entre otras. La ortogonalidad de las bases y la linealidad de la fase de los filtros son propiedades incompatibles, exepto para la *wavelet* Haar [55], pero se pueden usar *wavelets* bi-ortogonales. La principal atracción de las *wavelets* bi-ortogonales es la fase lineal de los filtros FIR. La fase lineal en los filtros hacen fácil la implementación del algoritmo piramidal sin ser necesaria la compensación de fase [79], [4].
En [45], se compara el desempeño relativo de las *wavelets* de fase lineal y las *wavelets* de fase mínima para la detección del periodo del pitch usando un algoritmo de detección de

eventos basado en la DyWT. En este caso se aplica la DyWT para detectar el instante de cierre glótico en las señales de voz.

Localización en el plano tiempo-frecuencia

Una de las motivaciones principales por las cuales la transformada *wavelet* es ampliamente usada es su capacidad de proveer representaciones en tiempo-frecuencia de señales, con buenas propiedades de localización en ambas variables. Se sabe que si ψ está bien localizada en tiempo y frecuencia, entonces el marco generado por ψ tendrá esa misma propiedad [24]. Si ψ está bien localiza en tiempo y frecuencia, entonces $\psi_{j,n}$ estará bien localizada alrededor $s_0^j nu_0$ en tiempo y alrededor de $s_0^{-j} w_0$ en frecuencia.
A partir de la experiencia se ha encontrado que una representación en tiempo-frecuencia $\rho_f(t, f)$ de alguna señal $f(t)$ debe cumplir las siguientes propiedades [14]:

- La distribución de tiempo-frecuencia (TFD, Time Frecuency Distribution) debería ser real, y su integración sobre todo el dominio de tiempo-frecuencia debería ser igual a la energía de la señal $f(t)$.

- La energía de la señal en una determinada región R del plano (t, w), debería ser igual a la obtenida por integración de $\rho_f(t, w)$ sobre los límites de $(\Delta t, \Delta w)$ de la región R:

$$\int_{\Delta t} \int_{\Delta w} \rho_f(t, w) dw dt = E_{fr} \tag{3.34}$$

  donde E_{fr} es un porción de la energía de la señal en la banda de frecuencia Δw y en el intervalo de tiempo Δt.

- El pico de la representación de tiempo-frecuencia una señal de frecuencia modulada monocomponente debería reflejar la frecuencia intermedia de la señal:

$$\left. \frac{\partial \rho_f(t, w)}{\partial w} \right|_{w=w_i(t)} = 0 \tag{3.35}$$

- Se espera que la TFD tenga buena concentración de energía alrededor de la frecuencia intermedia para una señal de FM.

Capítulo 4

Selección de bases

Las señales de voz son un tipo de señales complejas en las cuales se encuentran variados tipos de estructuras que cambian tanto en tiempo como en frecuencia, debido a ello podría ser más conveniente poder representar una señal dependiendo las estructuras que la conforman.

Al trabajar con *wavelets* se tiene la posibilidad de escoger en particular las bases *wavelet* que convengan para la tarea que se realiza. La escogencia puede hacerse dentro de un diccionario de bases ó se puede escoger la *wavelet* madre usada para la WT. La selección de la *wavelet* básica o *wavelet* madre depende mucho de la naturaleza de la señal y de el objetivo del procesamiento digital de esas señales [78].

4.1. Wavelet packets

Las transformada *wavelet packet* (WP) es una generalización de la transformada *wavelet* discreta. La transformada *wavelet packet* es obtenida mediante el algoritmo de multiresolución descrito en el apéndice **A**, pero con la diferencia de que tanto los coeficientes de aproximación como los de detalle son descompuestos para cada nivel, procedimiento que genera una estructura de árbol [8], tal como se muestra en la figura (4.1).

Un espacio V_j de aproximación de multi-resolución es descompuesto es un espacio bajo de multi-resolución V_{j+1} sumado a un espacio de detalle W_{j+1}. Ello se hace dividiendo las bases ortogonales $\{\varphi_j(t-2^j n)\}_{n\in\mathbb{Z}}$ de V_j en dos nuevas bases ortogonales,

$$\{\varphi_{j+1}(t-2^{j+1}n)\}_{n\in\mathbb{Z}} \quad de \quad V_{j+1} \quad y \quad \{\psi_{j+1}(t-2^{j+1}n)\}_{n\in\mathbb{Z}} \quad de \quad W_{j+1} \qquad (4.1)$$

En lugar de dividir únicamente los espacios de aproximación V_j para construir los espacios de detalle W_j y las bases *wavelet*, lo que se hace es dividir estos espacios de detalle para obtener nuevas bases. Algún nodo del árbol binario es etiquetado por su nivel de descomposición j y el número p de nodos que están a su izquierda en el nivel de descomposición j, como se ilustra en al figura (4.1). Por inducción se construye un árbol binario donde cada nodo (j, p) corresponde a un espacio W_j^p. En la raíz del árbol se encuentra $W_0^0 = V_0$. Este espacio de aproximación admite una base ortogonal de funciones de escalamiento $\{\varphi(t-n)\}_{n\in\mathbb{Z}}$, de tal forma que $\psi_0^0 = \varphi$. Supóngase ahora que se ha construido W_j^p y sus bases ortonormales $\mathcal{B}_j^p = \{\psi_j^p(t-2^j n)\}_{n\in\mathbb{Z}}$ en el nodo (j, p) [55]. La dos bases ortogonales en los nodos hijos son definidos por las relaciones,

$$\psi_{j+1}^{2p}(t) = \sum_{n=-\infty}^{\infty} h[n]\psi_j^p(t-2^j n) \tag{4.2}$$

y

$$\psi_{j+1}^{2p+1}(t) = \sum_{n=-\infty}^{\infty} g[n]\psi_j^p(t-2^j n) \tag{4.3}$$

En [88] se prueba que $\mathcal{B}_{j+1}^{2p} = \{\psi_{j+1}^{2p}(t-2^{j+1}n)\}_{n\in\mathbb{Z}}$ y $\mathcal{B}_{j+1}^{2p+1} = \{\psi_{j+1}^{2p+1}(t-2^{j+1}n)\}_{n\in\mathbb{Z}}$ son bases ortonormales de dos espacios ortogonales W_{j+1}^{2p} y W_{j+1}^{2p+1} tal que,

$$W_{j+1}^{2p} \oplus W_{j+1}^{2p+1} = W_j^p \tag{4.4}$$

Esta división recursiva define un árbol binario de espacios de paquetes *wavelet* donde cada nodo padre es dividido en dos subespacios ortogonales.

Para cada etapa en la descomposición, el algoritmo *wavelet packet* particiona el plano tiempo-frecuencia en rectángulos de forma constante. Ellos se hacen más amplios (en tiempo) y más angostos (en frecuencia) a medida que la descomposición avanza en nivel. Esto es mostrado esquemáticamente en la figura (4.1) para tres escalas o tres niveles de descomposición que es lo mismo. A medida que el nivel de descomposición avanza se obtienen particiones del plano tiempo-frecuencia de forma diferente. Cada nivel de descomposición entrega una partición del plano tiempo-frecuencia que al unirlo con las particiones de los

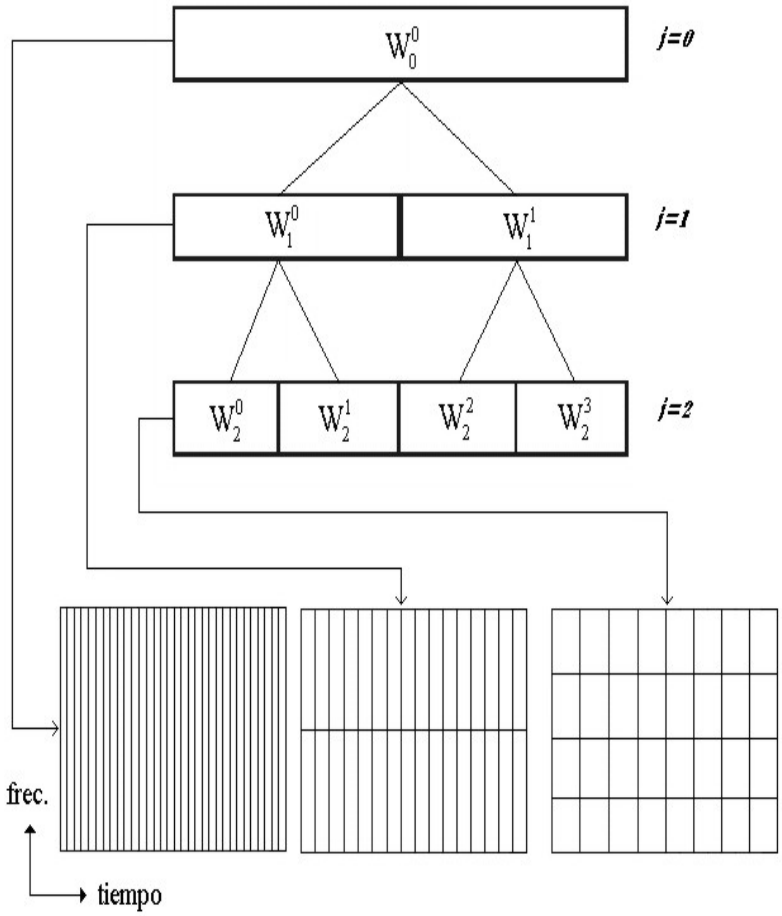

Figura 4.1: Diagrama esquemático de la descomposición en paquetes *wavelet*

otros niveles de descomposición forman una gran familia de particiones del plano tiempo-frecuencia. Así, se mejora enormemente las opciones de escogencia para los algoritmos de selección de bases de representación [21] y clasificación [59].

4.1.1. Funciones de costo

Antes de seleccionar la mejor base (base óptima), es importante introducir el concepto de costo de información, o el costo de almacenar una representación escogida.
Pueden ser definidas diversas funciones de costo de información M, las más usadas son aquellas que miden concentración. Esto significa que M tiene que ser grande cuando los

elementos de la secuencia son básicamente del mismo tamaño, y pequeña cuando todos, o al menos algunos, son despreciables. Algunos ejemplos de funciones de costo de información son [64], [88]:

- **Umbral**: Esta función de costo de información consiste en escoger un limitador arbitrario ϵ. Establece el número de elementos de una secuencia necesarios para transmitir la señal de tal forma que se reciba con una precisión ϵ.

- **Entropía**: Se define la entropía de Shannon para una distribución de probabilidad discreta q_k para $k = 1, 2, ..., N$, de la forma

$$\mathcal{H}(q) = \sum_k q(k) \log \frac{1}{q(k)} \tag{4.5}$$

  La máxima entropía posible se consigue para una función de probabilidad equiprobable, es decir, cuando la información está esparcida a través de todo el vector. Alguna otra distribución resulta en un valor menor de entropía. La mínima entropía ocurre cuando toda información recae en una simple localización k [8]. Esta medida puede ser aplicada a los coeficientes *wavelet* de energía, donde se desea retener la mayor cantidad de información posible de la señal en unos pocos coeficientes. Para utilizar la entropía de Shannon en la selección de bases *wavelet* se trabaja con la energía normalizada de los coeficientes *wavelet*, donde $q(k)$ vendría a ser ahora dicha energía normalizada. En la figura 4.2 se ilustra el funcionamiento de la medida de entropía de Shannon, tomando a 10 como la base para el logaritmo.

4.1.2. Algoritmo de *best basis*

Sea B_j^p un conjunto de vectores base que pertenecen al subespacio W_j^p ordenados de la forma:

$$B_j^p = (w_{j,0}^p, ..., w_{j,2^{n_0-j}-1}^p) \tag{4.6}$$

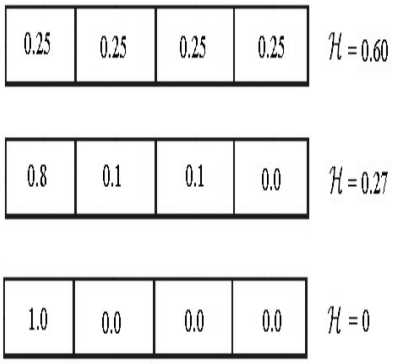

Figura 4.2: Funcionamiento de la entropía de Shannon

Donde 2^{n_0-j} es la cantidad de vectores base $\{w_{j,l}^p\}_{l=0}^{2^{n_0-j-1}}$ que realizan el cubrimiento para cada sub-espacio W_j^p. Además, sea A_j^p la mejor base para la señal f restringida al *span* de B_j^p y sea M una función de coste de información que mide la habilidad de los nodos(subespacios) para la compresión. El algoritmo se desarrolla como sigue:

- Escoja un diccionario de bases ortonormales $\mathcal{D}$
- Expanda la señal f en el diccionario $\mathcal{D}$ y obtenga los coeficientes $\{B_j^p f\}_{0\leq j\leq J, 0\leq k\leq 2^j-1}$
- Haga $A_J^p = B_J^p$ para $k = 0, ..., 2^J - 1$
- Determine el mejor subespacio A_j^p para $j = J-1, ..., 0\, p = 0, ..., 2^j - 1$, de la forma:

$$A_j^p = \begin{cases} B_j^p & si \;\; M(f, A_j^p) \leq M(f, A_{j+1}^{2p}) + M(f, A_{j+1}^{2p+1}) \\ A_{j+1}^{2p} \oplus A_{j+1}^{2p+1} & otros \;\; casos \end{cases} \tag{4.7}$$

El conjunto de coeficientes de la transformada WP, que poseen menor entropía, son seleccionados para representar la señal. Es decir, se desea que la información de la señal sea concentrada en la menor cantidad de coeficientes posible. Para cada escala, cada conjunto de coeficientes particionados de a pares, los hijos, son comparados con aquel conjunto de coeficientes del cual provienen, el padre. Si la suma de las entropías de los hijos es menor que la entropía del padre, entonces dicho hijos se mantienen (No son eliminados). Esto es

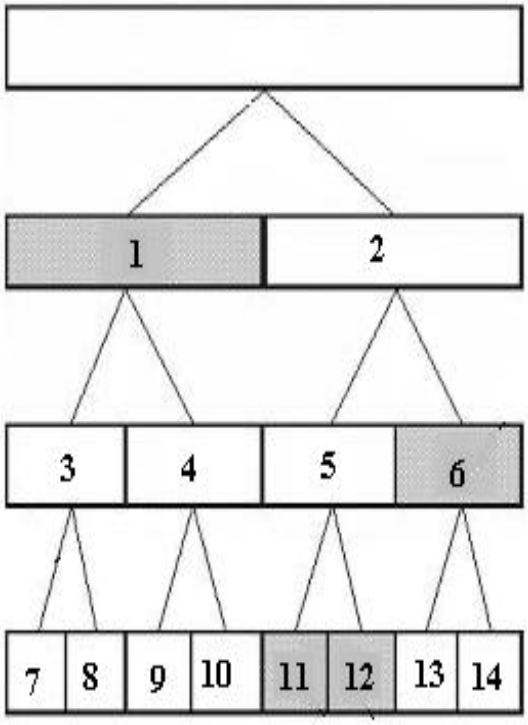

Figura 4.3: Algoritmo de *best basis* usando la entropía como función de costo, [8]

mostrado esquemáticamente en la figura 4.3. Dado que la entropía de los nodos hijo 3 y 4 es mayor que la entropía de su nodo padre, no son seleccionados. La entropía combinada de los nodos 11 y 12 es menor que la de su padre, por ello, son seleccionados. La entropía del subespacio 6 es menor que la entropía combinada de los nodos 13 y 14 por lo tanto, es seleccionado.

4.2. Selección de la wavelet madre para la estimación del pitch

Las *wavelet* madre usadas en la literatura para la estimación del pitch generalmente corresponden a la *spline* de orden 2 y 3, a la *Daubechies* de orden 1 y a la Morlet o gaussiana modulada, pero en dichos trabajos no se deja en claro los motivos por los cuales se usó determinada *wavelet* madre.

En la estimación de la frecuencia fundamental la determinación de los máximos locales juega un papel importante; el ideal es obtener una representación en la cual, los únicos máximos locales que existan correspondan a GCI's, pero tal hecho no sucede [16]. En este trabajo, se plantea una metodología para la selección de la *wavelet* madre ψ en la representación de señales de voz, orientada a la estimación del pitch. La selección debe ser tal, que produzca la máxima diferenciación entre los valores de los coeficientes de representación asociados al CGI y los demás.

En (3.2.4) se muestra que si f es regular y ψ tiene una cantidad *p* suficiente de momentos de desvanecimiento, entonces, los respectivos coeficientes WT $\mid \langle f, \psi_{j,n} \rangle \mid$ serán pequeños a escalas finas de 2^j.

Si f presenta una singularidad aislada en el momento t_0, el cual se encuentra dentro del soporte compacto de $\psi_{j,n}(t) = \frac{1}{\sqrt{2}}\psi(\frac{t-2^j n}{2^j})$, entonces, el producto $\langle f, \psi_{j,n} \rangle$, en forma general, le corresponderá una amplitud grande. Si ψ tiene soporte compacto de longitud K, en cada escala 2^j existirán K funciones *wavelet* $\psi_{j,n}$, cuyo soporte incluye a t_0. Para minimizar el número de coeficientes de amplitud grande se debe reducir el tamaño del soporte de ψ [55].

En el caso en que f tenga pocas singularidades aisladas y sea suficientemente suave entre dichas singularidades, es preferible el uso de *wavelets* madre con bastantes momentos de desvanecimiento, en orden a obtener la mayor cantidad de coeficientes $\langle f, \psi_{j,n} \rangle$ de valor cercano a cero. En cambio, si la densidad de singularidades por unidad de tiempo se incrementa, sería recomendable reducir el tamaño del soporte, aunque su costo sería la reducción de los momentos de desvanecimiento.

Respecto al compromiso entre la menor longitud del soporte compacto contra la mayor cantidad de los momentos de desvanecimiento, las *wavelets* que mejor se desempeñan son las pertenecientes a la familia *Daubechies* y las del tipo *Spline* [55].

En [45] se compara el desempeño relativo de las *wavelets* de fase lineal y las *wavelets* de fase no lineal para la detección de eventos del *pitch* usando un algoritmo de detección de eventos basado en la DyWT. Empleada para detectar el cierre glótico. En dicho trabajo se reporta que las *wavelets* de la familia *spline* entregan mejores resultados.

Cuando las características de la función *wavelet* usadas para determinar la *wavelet* madre no son suficientes, entonces la medida de entropía puede ser utilizada para la selección de

la *wavelet* más conveniente, la función de entropía de Shannon puede ser utilizada para tal fin [8].

4.3. Selección de bases discriminantes

En ésta sección se describe un algoritmo rápido para construir extractores de características. Se seleccionan bases que están bien localizadas en el plano tiempo-frecuencia de tal forma que sean discriminantes para determinadas clases pre-establecidas.

4.3.1. Medidas discriminates

En la sección 4.1.1 se expusieron varias funciones de costo, entre ellas la más destacada la función de entropía de Shannon; ésta cantidad mide la variabilidad de la magnitud de los coeficientes y fue usada en el algoritmo de selección de bases propuesto por Coiffman y Wikerhauser [21]. La entropía mide la eficiencia de cada subespacio para representar la señal, lo cual es útil desde el punto de vista de compresión (representación). Pero, en problemas de clasificación, lo que se desea es medir el poder discriminante de cada subespacio que pertenece a las bases del árbol obtenido a través de la transformada WP en lugar de medir la eficiencia de la representación. Lo primero que se hace es seleccionar la medida discriminante, y una vez se selecciona, se compara la bondad de cada nodo(subespacio), para problema de clasificación, respecto a la brindada por sus nodos hijos y de esa forma establecer si es o no conveniente realizar la descomposición del nodo(subespacio) en sus nodos hijos [59].

Existen varias opciones entre las cuales escoger la medida discriminante, [73], pero en esencia lo que ellas hacen es medir la distancia estadística entre clases. Por simplicidad, considérense dos clases. Sea $\mathbf{p} = \{p_i\}_{i=1}^n$, $\mathbf{q} = \{q_i\}_{i=1}^n$ dos secuencias no negativas para las que se cumple que $\sum p_i = \sum q_i = 1$ (las cuales pueden ser vistas como distribuciones de energía normalizadas pertenecientes a las clases 1 y 2 respectivamente). La función de costo de información discriminante $\mathcal{D}(\mathbf{p}, \mathbf{q})$ entre estas dos secuencias debería medir que tan diferente están distribuidos los vectores $\mathbf{p}$ y $\mathbf{q}$. Una búsqueda natural es la *entropía*

relativa también conocida como *distancia de Kullback-Lieber*, definida por [50]:

$$\mathbf{I}(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \tag{4.8}$$

Esta cantidad no es una distancia(métrica), básicamente porque no es simétrica. Una versión simétrica la da la divergencia $\mathbf{J}$ entre $\mathbf{p}$ y $\mathbf{q}$, también denominada divergencia de Kullback, la cual está definida por:

$$\mathbf{J}(\mathbf{p},\mathbf{q}) = \mathbf{I}(\mathbf{p},\mathbf{q}) + I(\mathbf{q},\mathbf{p}) \tag{4.9}$$

Otra posible función para $\mathcal{D}$ sería

$$W(\mathbf{p},\mathbf{q}) = \| \mathbf{p} - \mathbf{q} \|^2 = \sum_{i=1}^{n} (p_i - q_i)^2 \tag{4.10}$$

Para medir la discrepancia entre una cantidad C de distribuciones, $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}$,..., $\mathbf{p}^{(C)}$, se toma la medida de discrepancia entre pares, y la medida final estaría dada por:

$$\mathcal{D}(\{\mathbf{p}^{(c)}\}_{i=1}^{C}) = \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} D(\mathbf{p}^{(i)}, \mathbf{p}^{(j)}) \tag{4.11}$$

4.3.2. El algoritmo Local Discriminant Bases

Recientemente los autores han introducido el concepto de *Local Discriminant Basis* [71, 69,59,68] en problemas de clasificación de señales e imágenes; entre ellas señales acústicas de origen geofísico [70, 72].

En esta sección se explora la extracción de características relevantes, a partir de la transformada WP de señales vía clasificación. El algoritmo propuesto en [59] es un algoritmo que de forma rápida selecciona las bases eficientes, a partir de un diccionario de bases ortonormales, que mejora el desempeño del esquema de clasificación en particular.

El primer paso en este algoritmo consiste en calcular los mapas de energía de tiempo-frecuencia para cada clase, de la forma:

Definición 4.1. *Sea* $\{f_i^{(c)}\}_{i=1}^{N_c}$ *un subconjunto de señales de entrenamiento que pertenecen a la clase* c*. Entonces el mapa de tiempo-frecuencia de la clase* c*, denotado por* Γ_c*, es una matriz de valores reales especificados por al tripleta* (j, p, l) *dada por:*

$$\Gamma_c(j,p,l) = \sum_{i=1}^{N_c} \frac{(w_{j,l}^p f_i^{(c)})^2}{\sum_{i=1}^{N_c} \parallel f_i^{(c)} \parallel^2} \tag{4.12}$$

para $j = 0, ..., J$*,* $p = 0, ..., 2^j - 1$*,* $l = 0, ..., 2^{n_0 - j} - 1$*.*

En otras palabras, γ_c se obtiene por la acumulación sucesiva de los coeficientes de expansión al cuadrado de las señales para cada posición en la matriz seguida de la normalización por la energía de las señales pertenecientes a la clase c.

Una vez las energías de las bases son acumuladas y normalizadas, para cada clase separadamente, se forma una distribución de energía de tiempo-frecuencia por clase. Basados en la diferencia entre estas distribuciones de energía (medida por alguna función de costo) un conjunto completo de bases ortonormales es seleccionado. Luego las características más relevantes alimentan algún clasificador en particular.

Capítulo 5

Estimación de características de voz usando transformada wavelet

La extracción de características relevantes de señales es un importante proceso en el análisis de datos, entre ellas la clasificación de señales en diferentes categorías. La manipulación directa del espacio de las señales para clasificación es totalmente impráctico debido a que: 1) el espacio de señales generalmente es de alta dimensionalidad, y 2) la existencia de ruido o componentes indeseados hacen que la clasificación sea difícil. De otra parte, el espacio de señales es altamente redundante comparado con el espacio de características. Además es importante reducir la dimensionalidad del problema, es decir, extraer sólo las características relevantes para el problema en particular y descartar información irrelevante. Es este sentido se crea un *espacio de características* $\mathcal{F} \subset \mathbb{R}^k$ donde $k \leq m$ y m es la dimensionalidad del espacio de señales o la longitud de cada señal. Un *extractor de características* se define como el mapeo $g : \mathcal{X} \longrightarrow \mathcal{F}$ [68].

En el análisis de la voz es común el empleo de dos tipos de características: las *características acústicas* (CA), que califican las cualidades vocales y poseen un sentido físico determinado; y las *características de representación*, que corresponden a valores calculados a partir de alguna forma de representación de la voz, y a los cuales, en general, no les corresponde algún sentido físico [83]. El uso de características acústicas en la descripción de las cualidades patológicas de la voz ha sido probado en varios contextos y con una variedad de objetivos. Uno de sus atractivos se centra en el hecho de que podría entregar una evaluación cuantitativa de las características vocales que de otra forma sería difícil medir [23].

Los parámetros acústicos se pueden clasificar de la forma,

- *Parámetros Cuasiperiódicos*, a los cuales pertenecen las características: Frecuencia fundamental, Formantes.
- *Parámetros de perturbación*: Jitter, Shimmer, Relación Armónico-Ruido, Relación excitación glótica - ruido.

El uso de características basadas en la transformada *wavelet* ha probado ser muy efectivo en tareas de clasificación [65, 1] y de estimación de determinadas características acústicas [44, 87, 28]. Las características de representación se encargan de describir el comportamiento dinámico de señales, las cuales son tomadas a partir de métodos de representación de señales (LPC (*Linear Predictive Coding*), cepstrum,...), y generalmente no se les asocia un sentido físico.

5.1. Segmentación de los tipos de voz sonoro/sordo

La clasificación sonoro/sordo consiste en determinar si está o no involucrada en el sistema de producción de la voz la vibración de la cuerdas vocales [38]. La clasificación de los segmentos cortos de voz en sonoro ó sordo es vital en los sistemas de análisis y síntesis y recientemente se ha probado la efectividad de la WT para tal fin [43, 74, 40].
La segmentación sonora/sorda se puede clasificar en dos tipos, por detección de eventos (segmentación no uniforme) y por determinación de ventana fija (segmentación uniforme). En esta última, se escoge un tamaño de ventana fijo para luego determinar si cada ventana pertenece a uno u otro tipo de voz. En el presente trabajo se comparan dos métodos de segmentación por ventanas fijas, el primero de ellos se realiza basado en análisis LPC (Linear Predictive Coding) implementado dentro del *toolbox* desarrollado por D. G. Childers en [19]; una explicación detallada de dicho método se explica en el anexo **B**.
Existe varios métodos para la segmentación de los tipos sonoro y sordo usando WT. Y una buena cantidad de ellos realiza la segmentación al mismo tiempo que realiza la estimación de la frecuencia fundamental. En [61] las trayectorias de máxima amplitud de la

WT causadas por los GCI son usadas tanto para la determinación del GCI como para la segmentación. Dado que las TMA (Trayectorias de máxima amplitud) son más organizadas y por ende de mayor energía para los sonidos del tipo sonoro que para los del tipo sordo, lo que se hace es verificar la energía de las trayectorias en cada ventana de análisis respecto a un umbral.

En [28], para la estimación del *pitch* se establecen máximos locales correspondientes a GCI's. Para realizar la segmentación no uniforme cuentan los máximos consecutivos correspondientes a instantes de cierre glótico; de existir una cantidad mayor a M máximos consecutivos, se deduce que tal porción de señal pertenece al tipo sonoro.

La energía para los sonidos del tipo sonoro se ubican en las escalas correspondientes a las bajas frecuencias, mientras que para los sonidos del tipo sordo, entre ellos los fricativos, la energía tiene a ubicarse en las escalas correspondientes a las frecuencias altas. El algoritmo usado en el presente trabajo aprovecha está propiedad. La energía de la escala más pequeña (altas frecuencias) debe ser mayor que la energía contenida en las otras escalas para que la ventana de análisis sea clasificada como sorda.

5.2. Frecuencia Fundamental

El *pitch* o *frecuencia fundamental* F_0 se determina por la velocidad de apertura o cierre de las cuerdas vocales en la laringe durante la fonación de sonidos sonoros, cuyo inverso corresponde a su vez al *periodo fundamental* T_0. La estimación del pitch es importante en aplicaciones como la codificación de voz, el desarrollo de sistemas de ayuda a discapacitados (entrenamiento de sordos) [28]. El pitch se emplea en la determinación de la entonación y las características emocionales de la voz. Así mismo, sus desviaciones pueden indicar la presencia de desórdenes funcionales y patologías [19].

En forma general, la definición de periodicidad se determina para intervalos infinitos de análisis, sin embargo, para efectos prácticos (por ejemplo de cálculo computacional), su estimación se realiza sobre intervalos finitos, que permitan cubrir varios periodos del pitch, o de manera instantánea a partir de la diferencia entre dos momentos consecutivos del cierre glótico [83].

En la estimación del pitch, se consideran las siguientes restricciones [9]:

- Los segmentos de voz son altamente no estacionarios, o corresponden al producto de vibraciones irregulares de las cuerdas vocales.
- La exitación glótica no es rigurosamente periódica.
- Existe una importante interacción entre la exitación y el tracto vocal, en donde la periodicidad que se observa en la señal resultante, es debida a la acción conjunta de la exitación casi-periódica y los primeros formantes de banda estrecha.
- La segmentación del inicio y del final de las zonas sonoras debe realizarse con alta precisión, tarea que no siempre es fácil de implementar.
- Alto margen dinámico de variación del parámetro F_0. Por tanto, dicha estimación debe restringirse a un intervalo de valores permitidos, en donde las ventanas de tiempo se ajusten dependiendo del hablante considerado, pudiendo abarcar entre 2 y más de $20\ ms\ (50 - 500Hz)$, para cubrir voces desde niños o sopranos, hasta barítonos, o entre 6 y 12 ms (80-170Hz), para el caso de locutores adultos promedio.

5.2.1. Estimación de la frecuencia fundamental basada en trayectorias de máxima amplitud

Las técnicas más comunes para la determinación de la frecuencia fundamental usando WT se basan en la propiedad que tiene la WT de producir máximos locales en puntos de singularidad de la señal, tal como se hizo en [44, 58]. Otras técnicas para la estimación de la frecuencia fundamental se basan en el cálculo de la transformada *Wavelet* continua [28,91], en [28] se usa la función Morlet a modo de *Wavelet madre*. La frecuencia fundamental aparecerá entonces como una línea horizontal en la representación tiempo-escala. Debido a la localización limitada de la WT en el dominio de la frecuencia; la frecuencia fundamental y los formantes aparecerán a modo de bandas esparcidas [16]. En [41, 28], se aprovecha la propiedad de buena resolución conjunta en tiempo-frecuencia de la WT para la localización de cambios abruptos que ocurren en los instantes de cierre glótico.

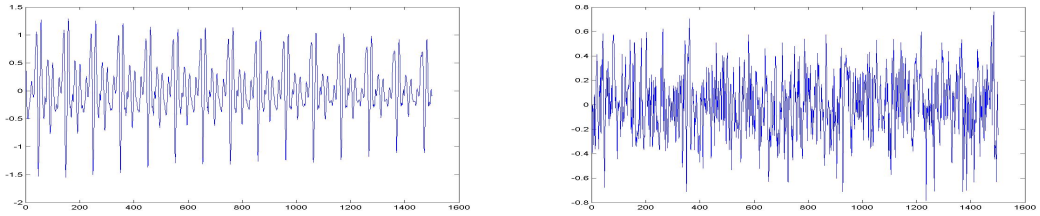

Figura 5.1: Escala de mayor detalle para una señal sin ruido y con ruido, respectivamente.

Los algoritmos mostrados en [52] y [61] construyen trayectorias de máxima amplitud a través de las escalas de descomposición de la WT, para luego determinar cuales de ellas corresponden a GCI. Aquellas escalas asociadas a las frecuencias altas poseen mayor resolución en el tiempo, por lo cual se usan sus máximos para determinar de manera más exacta el GCI.

En [52] se reporta éxito en la determinación de los GCI, el cual se basa en un algoritmo de programación dinámica. Las escalas de la WT asociadas a las altas frecuencias, poseen mayor resolución en el tiempo, por lo cual usa los máximos locales ubicados en dichas escalas para determinar adecuadamente el GCI. En [28, 61, 52], no se deja en claro el comportamiento de los algoritmos ante el ruido. Pero debido al hecho de que las escalas asociadas a las altas frecuencias sufren más los efectos del ruido aleatorio, ver figura 5.1, es de esperarse que los algoritmos mostrados en [52] y [61] bajen su rendimiento ante el ruido.

Determinación de las trayectorias de máxima amplitud Los coeficientes de máxima amplitud de la WT pueden ser calculados para cada escala. Estos valores máximos son definidos como cualquier punto t_0 en la escala i, tal que $y_i(t_0) > y_i(t)$, donde t pertenece a la vecindad derecha o izquierda de t_0. Seguidamente, se organizan los máximos locales en trayectorias de máxima amplitud (TMA), los cuales deberían estar conectados formando trayectorias rectilíneas. Pero debido a las diferencias en fase de la señal para diferentes escalas, las líneas no son rectas.

Sea $M_a(i, j)$ la *i*-ésima amplitud máxima en la escala j. Además, sea $L_m(i, j)$ la TMA, cuya búsqueda se empieza por la escala de frecuencia de menor a mayor. Las amplitudes

acumuladas $A_c(i,j)$ a lo largo de las trayectorias son calculadas usando las siguientes ecuaciones locales [61]:

$$A_c(i,j) = \text{máx} \begin{cases} \frac{A_c(il,j-1)}{lw} + M_a(i,j) \\ A_c(i,j-1) + M_a(i,j) \\ \frac{A_c(ir,j-1)}{rw} + M_a(i,j) \end{cases} \tag{5.1}$$

donde il, ir son los índices de la máxima amplitud en la respectiva vecindad (izquierda,derecha) de $M_a(i,j)$ sobre la escala $j-1$, siendo lw (rw) el factor de peso del valor absoluto de la diferencia entre i e il (ir), por ejemplo, $lw = i - il$ y $(rw = ir - i)$. Los índices il e ir son determinados comparando el peso de las amplitudes máximas en las respectivas vecindades de i, el cual es inversamente proporcional a la distancia al centro de frecuencia de la escala.

Determinación del instante de cierre glótico Toda singularidad en una señal produce máximos a través de varias escalas de la respectiva descomposición *wavelet*, mas no toda línea de máximos locales representa una singularidad asociada al algún GCI, generando problemas en su detección [55]. Por lo tanto, se debe distinguir entre las trayectorias que corresponden a instantes de GCI y las debidas a otra naturaleza como se ilustra en la figura 5.2, y su respectiva gráfica de trayectorias de máximos locales 5.3.

En el caso de una señal insonora las TMA tienen poca energía, debido a que las amplitudes máximas se encuentran a lo largo de las escalas de descomposición (plano tiempo-escala). En general, para escoger los GCI, se toman aquellas TMA de mayor energía [61]. Por cuanto, tal regla no es suficiente, en este trabajo se encontró que, de existir alguna TMA no asociada a la aparición del GCI, el valor de su energía es menor que el promedio, multiplicado por un factor de ajuste, de los valores respectivos de los TMA vecinos asociados al GCI. El algoritmo de determinación del GCI es el siguiente:

$$m_i = \rho(\zeta(i-1) + \zeta(i+1))/2 \tag{5.2}$$

$$\zeta(i) \Longleftrightarrow GCI \quad si \quad m_i < \zeta(i) \tag{5.3}$$

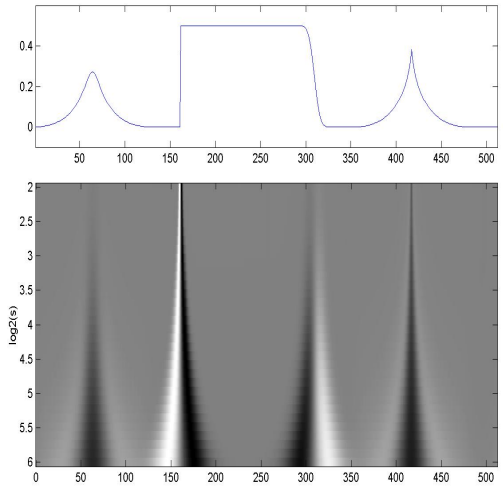

Figura 5.2: Transformada *wavelet* como función de u y $\log_2 s$

donde $\zeta(i)$ es la energía asociada a la *i*-ésima TMA y ρ es el factor de ajuste.
La segmentación se realiza comparando la energía de las trayectorias buscadas con un umbral θ, empíricamente establecido para los segmentos sonoros.

5.3. Parámetros de medida de perturbación

Una medida de perturbación es un valor efectivo de la perturbación global de una característica. Existe una variedad de medidas de perturbación, entre ellas la desviación estándar. Para la estimación de las perturbaciones de los parámetros de frecuencia fundamental y amplitud pico, es común el empleo del promedio relativo de perturbación (RAP Relative Average Perturbation) definido como [83].

$$\vartheta_{RAP} = \frac{\left(\sum_{i=1}^{n} \mid \frac{z_{i-1}+z_i z_{i+1}}{3} - z_i \mid\right)}{\sum_{i=1}^{n} z_i} \tag{5.4}$$

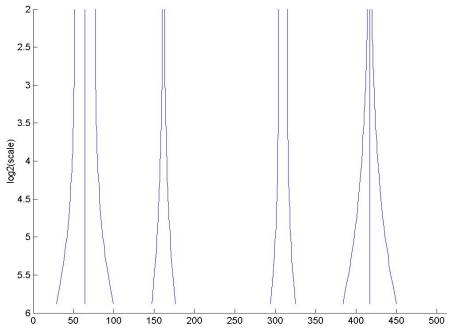

Figura 5.3: *Modulus maxima* de la transformada *wavelet*

Jitter

Se refiere a la perturbación en la frecuencia fundamental de la señal de voz. El cálculo del jitter requiere por tanto la determinación del pitch. En el caso de las voces patológicas, la envolvente en el tiempo de la señal, parece ser diferente a la envolvente de las voces normales. Estas variaciones fueron vistas por primera vez en [49]. Adicionalmente en las voces patológicas se ha encontrado la aparición de pulsos cíclicos que pueden perturbar la medida de la frecuencia fundamental [80], generando, una fuerte variación en la medida del Jitter. Se obtiene al reemplazar en 5.4 la variable z_i por la frecuencia fundamental.

Shimmer

Representa una medida de la perturbación de la amplitud del pico máximo de la señal pico a pico. Este parámetro sirve para cuantificar pequeños lapsos de inestabilidad vocal. Se obtiene al reemplazar en 5.4 la variable z_i por el valor pico a pico de la señal de voz.

5.4. Parámetros de ruido

Los parámetros basados en la relación entre la energía armónica y la energía de ruido tiene amplia aplicación en la clínica de la voz, por su estrecha relación con muchas disfonías [23], [37], [51]. Por otra parte son muy sensibles a los sistemas de registro empleados ya

que éstos pueden introducir niveles de ruido ajenos a los de la propia voz alterando los resultados. La formulación concreta de éstos resultados varía de unos algoritmos a otros. Por ejemplo, para calcular el HNR(razón armónico ruido, *harmonics-to-noise ratio*) o su recíproco NHR(razón ruido armónico, *noise-to-harmonics ratio*) en [37] se considera que la energía de ruido se encuentra en la región de los 2800 a los 5800 Hz, dado que la mayor parte del contenido de ruido de la voz pertenecen a altas frecuencias [37]. En [36] se toma la energía del ruido como aquel existente en la banda de los 1500 a los 4500 Hz.

Varios métodos se han propuesto para separar la señal de voz en sus componentes periódico y aperiódico: ellos están basados en modelado senoidal, en modelado armónico mas ruido, entre otros. Toda la energía presente en las frecuencias armónicas están asociadas a la componente determinística, pero en las señales de voz la energía del ruido están esparcido por todo el ancho de banda [11]. En [11] se obtiene la parte aperiódica de la señal basado en un procedimiento iterativo usando las técnicas de cepstrum y LPC(*Linear Predictive Coding*) combinadas. En [2] se usa un procedimiento de de-noising basado en umbralización de los coeficientes *wavelet* para la separación del ruido de la señal de voz, para luego obtener la relación señal a ruido de la voz, el procedimiento usado allí es especialmente útil para los casos en que el ruido es blanco gaussiano. Para la detección de periodicidades en series de tiempo a partir de la CWT lo que se hace es construir un esqueleto sobre el plano tiempo-frecuencia manteniendo solo aquellos componentes *wavelet* que son máximos locales respecto a las escalas para cada instante de tiempo, pero descartando del análisis aquellos que pertenecen a singularidades de la señal, tal como se hace en [39]. Tal método es costoso en términos de cómputo, además de encontrarse problemas al tratar de establecer trayectorias de máximos respecto a las escalas.

No en todos los problemas relacionados con extracción de ruido se tiene una definición exacta de los que es ruido, en lugar de ello en [22], Coifman y Wickerhauser proponen un método de reducción de ruido el cual tiene la habilidad de reducir el ruido incoherente en una señal sin afectar los componentes coherentes. La parte incoherente de la señal(ruido) es aquella que no está lo suficientemente correlacionada con las formas de onda de las bases. En la sección 4.1.2 se obtuvieron un conjunto de bases que mejor representan determinada señal a partir de un diccionario de bases. Se usa el algoritmo de *best basis selection* para

realizar la extracción de la estructura coherente de las señales de voz.

5.5. Parámetros de representación

Una señal $f \in \mathbf{H}$ es aproximada con M vectores seleccionados adaptivamente en una base ortonormal $\mathcal{B} = \{g_m\}_{m\in\mathbb{Z}}$ de $\mathbf{H}$. Sea f_M la proyección de f sobre M vectores cuyos índices están en I_M:

$$f_M = \sum_{m\in I_M} \langle f, g_m\rangle g_m \tag{5.5}$$

El error de aproximación corresponde a la suma,

$$\epsilon[M] = \| f - f_M \|^2 = \sum_{m\notin I_M} | \langle f, g_m\rangle |^2 \tag{5.6}$$

Para minimizar este error, los índices en I_M deben corresponder a los M vectores que tienen el valor del producto interno $| \langle f, g_m\rangle |$ más grande. Estos son los vectores que mejor correlacionan a f. Ellos pueden ser interpretados como las características principales de f [55].

Existen varios tipos de transformadas *wavelet*, donde el uso de cada una de ellas depende de la aplicación. La FWT es recomendada para trabajos de compresión y *denosing*, mientras que aquellas invariantes a las traslaciones en el tiempo son preferidas para propósitos de reconocimiento de patrones.

A pesar de que la FWT es no invariante a las traslaciones, dicha transformada ha sido usada con éxito para propósitos de clasificación en [62], [65]. La DyWT ha sido usada para propósitos de clasificación de voz en [40] y la transformada WP ha sido exitosamente usada en trabajos de reconocimiento en [59], [68], [6].

En [62] las características se seleccionan tomando los p coeficientes de mayor magnitud para cada nivel de aproximación de un total de 6 niveles de descomposición. El algoritmo implementado en dicho trabajo es una modificación del presentado en [12]. Se usaron 6 niveles de descomposición dado que después del sexto nivel de aproximación no se retiene información representativa del segmento de voz, es decir, que el aporte del séptimo nivel de aproximación tenía poco peso en el total de la señal respecto de los demás niveles.

La *wavelet* madre fue seleccionada probando las familias disponibles del *toolbox* de *Matlab* (*Daubechies, Symlets, Coiflets, Meyer, biorthogonal spline y reverse biorthogonal spline*). Se escogió aquella que entregase mayor porcentaje de clasificación respecto a los clasificadores Bayesiano, Redes Neuronales y máquinas de soporte vectorial, dando como resultado la *wavelet* de la familia Daubechies de orden 8. La base de datos usada estuvo conformada por voces de personas adultas, entre 19 y 54 años, hombres y mujeres, pertenecientes a las clases normal y disfónicas.

Además de usar los coeficientes a modo de características también se puede usar parámetros derivados de los coeficientes. En [6] se usa una combinación de la transformada *Wavelet Packet* y las redes neuronales para lograr la clasificación de las clases de voces disfónicas y normales. Además demuestran la viabilidad de la WP y el algoritmo de *Best Basis Selection*, expuesto por primera vez en [21], como extractor de características para sistemas de clasificación de voces disfónicas. En [64] se muestra que la función de costo más conveniente para el algoritmo de *Best Basis* es la entropía de Shannon y que la *wavelet* madre que entrega mejores resultados es la *Symlet* de orden 5. Al desarrollar el algoritmo de *Best Bases* a cada subesapcio de la descomposición WP se le asocia un valor de entropía, estos valores pueden caracterizar la señal que fue descompuesta, por ejemplo, si el nivel de descomposición es de 5, entonces los 63 valores del valor de entropía pueden caracterizar la señal [6].

Capítulo 6

Marco experimental

6.1. Base de datos fuente

En la prueba de los algoritmos desarrollados se determinó usar varias bases de datos. La primera de ellas es la base de datos *Keele Pitch Database*, del *Centre of Cognitive Neuroscience*, *The University of Liverpool* [57], con la cual se prueban los algoritmos de estimación del *pitch*. Adicionalmente, se usan otra bases de datos(la base de datos de voces disfónicas y la base de datos de personas con incompetencia velofaríangea) en las que se incluyen voces del tipo normal y patológico, todas ellas pertenecientes a la zona centro del país, recolectadas por el grupo de Control y Procesamiento Digital de Señales (GCyPDS).

6.1.1. Bases de datos de referencia

Para verificar el comportamiento en la estimación del *pitch*, el resultado obtenido debe confrontarse con equipos especializados o con métodos de calidad ya comprobada. Los equipos especializados son costosos y no siempre están disponibles en el medio, por ello, regularmente se prefiere la segunda opción. Pero afortunadamente se cuenta con una base de datos de referencia.

El detector basado en el cepstrum se ha utilizado ampliamente para tal fin [6]. Otra posible solución es la utilización de instrumentos de determinación del pitch, pero éste camino no siempre resulta sencillo debido a la resistencia de los locutores a usar estos instrumentos aparte de las modificaciones que pudiesen ser introducidas por los instrumentos en la forma de hablar. Una tercera solución es la determinación del pitch a partir de la forma de

onda de la señal entregada por el laringógrafo, dichas señales conforman nuestra base de datos para el establecimiento de una referencia. En el presente trabajo se usó un sistema semi-automático (su resultado se revisa manualmente) para la marcación de los instantes de cierre glótico. Para usar esta información en la evaluación de sistemas de estimación del pitch, a cada instante de cierre glótico se le asocia un valor del pitch equivalente a la distancia entre dicho instante y el anterior [28]. A partir de esta muestras se genera el contorno del pitch de referencia.

Los datos usados para el diseño experimental provienen de la base de datos Keele Pitch Database, del *Centre of Cognitive Neuroscience, The University of Liverpool*. Los datos corresponden a las salidas del laringógrafo y de voz tomadas simultáneamente para un texto preestablecido, el cual fue leído por 10 hablantes, 5 hombres y 5 mujeres [57].

En la base de datos, tanto la señal de voz como la del laringógrafo fueron tomadas a una frecuencia de muestreo de 22 kHz. De esta base de datos se usó un algoritmo de clasificación sonora/sorda para determinar las porciones de la señal a la cual se le extraería la frecuencia fundamental. El algoritmo de segmentación es una combinación del método usado en [19] y el método usado en [74]. A la señal proveniente del laringógrafo se le aplicó un algoritmo de marcación de instantes de GCI, se tomaron aquellos segmentos para los cuales se detectaron los GCI en un 100 %. A modo de segunda prueba, a los segmentos se les agregó ruido blanco gausiano, y de esa forma poder confrontar el funcionamiento del algoritmo en ambientes de ruido aditivo.

En la evaluación de la potencialidad del algoritmo propuesto se usa el esquema utilizado en [19] para la estimación de la frecuencia fundamental; el cual corresponde a una variante del algoritmo SIFT (*simplified inverse filtering tracking*). El SIFT busca la periodicidad de una señal estimada por filtrado inverso, y es uno de los métodos más comúnmente usados en equipos comerciales [28]. Los métodos de extracción del periodo fundamental por filtrado inverso tiene sus ventajas sobre los métodos basados en la autocorrelación y análisis cepstral. En la confrontación de los algoritmos se usa la medida de error cuadrático medio $error = \frac{1}{N}\sum_n^N (r(n) - p(t))^2$, donde $r(n)$ es el valor de referencia del *pitch* obtenido a partir de la base de datos *Keele Pitch Database* y $p(n)$ es la estimación del *pitch*. Para poder realizar la resta punto a punto entre $r(n)$ y $p(n)$ es necesario aplicar interpolación

Género	*Valoración*
42 Hombres	40 Normales
49 Mujeres	51 Disfonías

Tabla 6.1: Muestra de análisis y valoración del especialista para un total de 91 registros para cada vocal

por *splines* a $p(n)$, la cual pasa por todos los puntos a interpolar, y luego restarlo de $r(n)$.

6.1.2. Voces disfónicas

Es de interés reciente el análisis de señales de voz en la determinación patologías, por ello en el presente trabajo se incluye una base de datos de prueba de algoritmos de la cual hacen parte tanto voces normales como patológicas, todas ellas, voces de personas adultas.

- *Sujetos*. La muestra representativa de la población seleccionada, fue evaluada de forma subjetiva por el especialista en fonoaudiología, y una vez realizado el diagnóstico inicial se llevaron a cabo las respectivas sesiones de grabación con aquellas personas que amablemente colaboraron. Las voces recolectadas para ésta base de datos hacen parte de las clases: normal o disfónicas. La muestra de análisis para la presente base de datos consta de un total de 91 registros de voz, clasificados de la forma ilustrada en la tabla 6.1. De los 91 registros, 42 corresponden a voces del sexo masculino y 49 al sexo femenino. Del total de 91, 40 corresponden al tipo normal y 51 de ellas son voces disfónicas.

- *Recolección de señales de voz*. Las señales fueron tomadas bajo condiciones de bajo nivel de ruido ambiental usando un micrófono Shure SM58, el cual es dinámico unidireccional (cardiode) diseñado para vocalistas profesionales. Posee un filtro esférico incorporado que reduce los ruidos causados por el viento y el aliento. La dispersión polar de cardiode que posee, aisla la fuente sonora principal a la vez que reduce los ruidos de fondo. La distancia del micrófono al hablante está entre 10 y 15 cm.

Género	*Valoración*
16 Hombres	27 Normales
8 Mujeres	24 con LPH

Tabla 6.2: Muestra de análisis y valoración del especialista

Las propiedades de los archivos de audio generados para cada una de las señales son las siguientes:

- Formato: *.wav
- Frecuencia de muestreo: 22000Hz
- Bits por muestra: 16
- Canales: 1 (monofónico)
- La evaluación de las voces fue hecha por el especialista en la materia, el fonoaudiólogo.

6.1.3. Voces labio y/o paladar hendido

Las señales de voz fueron tomadas usando el micrófono SM58, el cual ya fue descrito en la sección anterior. Las clases dentro de las cuales fueron ubicadas las voces de los pacientes evaluados subjetivamente fueron: Normal o con LPH. Una vez hecha esta valoración, se procedía a la grabación de las voces.

– *Sujetos*. Las muestras corresponden a niños y niñas entre 5 y 15 años de edad, residentes de la zona centro del país. Se logro la grabación de un total de 51 voces, y tal como se muestra en la tabla 6.2 27 corresponden al tipo normal y 24 de ellas corresponde a pacientes con LPH

– *Recolección de voces*: Dicho procedimiento fue realizado en condiciones controladas de ruido ambiente. Las propiedades de los archivos de voz tienen las mismas propiedades descritas para los archivos de voces disfónicas. La evaluación de las voces fue hecha por el especialista en la materia, el fonoaudiólogo.

6.1.4. Conjunto de características a estimar

En el presente trabajo se presenta el análisis de las siguientes características: pertenecientes al tipo acústico y características de representación basadas en WT.

- Características acústicas: Frecuencia fundamental, medida de ruido de la señal de voz. El jitter pertenece es una medida de perturbación de la frecuencia fundamental y es de importancia reconocida, pero no se incluye en el presente trabajo por ser una medida tomada fácil y directamente del contorno del *pitch*, además, su importancia en la determinación de voces patológicas a ha sido ya ampliamente demostrada.
- Características de representación:
  - Coeficientes de DyWT.
  - El valor de la entropía para cada nodo de la WP fue usado a modo de características.
  - Coeficientes correspondientes a bases *wavelet* discriminantes.

6.2. Estimación del pitch

6.2.1. Selección de la Wavelet madre para la estimación del pitch

En la sección 4.2 se concluyó que las *wavelets* más convenientes, desde el punto de vista de singularidades, para la estimación del *pitch* corresponden a las de la familia *Spline* y *Daubechies*. Pero se requiere un método para seleccionar en concreto que *wavelet* madre usar. En el presente trabajo se propone usar la medida de entropía de Shannon para tal fin. Dentro de la familia de las *Spline* y *Daubechies*, en calidad de la mejor *wavelet* madre se escoge aquella que entregue la mayor cantidad de coeficientes cercanos a cero de tal forma que, sólo unos pocos sean de valor grande respecto de los demás. Con tal propósito, se usa la medida de variabilidad de la energía propuesta en [21], tomando el valor resultante al aplicar la función de entropía de Shannon calculada en cada escala de descomposición. La suma total de estos valores, por todas las escalas de descomposición, corresponde a la

función de costo final. De tal manera, que la selección final de la *wavelet* madre recae sobre la función que tenga el menor valor de la función de costo, estimada de la forma:

$$C = \min_{k} \sum_{\lambda=1}^{J} C_{k,\lambda}$$

donde $C_{k,\lambda}$ esta dado por

$$C_{k,\lambda} = -\sum_{m=1}^{N} \frac{|\langle f, \psi_{m,\lambda}\rangle|^2}{\|B^{\lambda}\|^2} \log_e \frac{|\langle f, \psi_{m,\lambda}\rangle|^2}{\|B^{\lambda}\|^2}$$

$\langle f, \psi_{m,\lambda}\rangle$ son los coeficientes en la posición m de la escala λ, y B^{λ} es la base *wavelet* asociada a la escala λ.

Respecto a la selección de la *wavelet* se usaron los registros pertenecientes a la base de datos de *voces disfónicas*, correspondientes a voces normales y voces patológicas, del tipo disfónico, entre hombres y mujeres. Todos los registro de voz corresponden a las vocales */a/, /e/, /i/, /o/, /u/*, las cuales pertenecen al tipo de voz sonoro. A pesar de haber seleccionado las de la familia *Spline* y *Daubechies* como las más convenientes para la estimación del *pitch* desde el punto de vista teórico, se incluyeron las familias del tipo *Coiflets* y *Symlets*. La cantidad de niveles de descomposición para la transformada *wavelet* diádica fue de 6.

A cada registro se le calculó la función de costo C dada por (**??**). Aquella *wavelet* para la cual dicha función fuese mínima se determinaba como la más conveniente para ese registro en particular. Lo mismo se hizo para todos los registros de la base de datos y se ordenaron por vocales. La cantidad de registros para los cuales determinada *wavelet* madre de prueba diese mejor fue entregado en porcentaje, tal como se muestra en la tabla 6.3.

Respecto a la vocal */a/*, según el criterio de función de entropía de Shannon, la *wavelet* madre *rbio 3.1*, la cual corresponde a la *Reverse Biortogonal*, que usa una *wavelet spline* de orden cúbico para la descomposición, y una de orden 1 para la reconstrucción, resulta ser la mejor para el $96{,}25\,\%$ de los registros, ver tabla 6.3. En la tabla 6.4 se observan aquellas *wavelets* que resultaron ser las segundas mejores según el criterio de variabilidad de la energía de Shannon.

Como se aprecia en la tabla, para la gran mayoría de los registros da que la wavelet rbio3.1, la cual corresponde a la spline de orden 3, es la más conveniente, la cual coincidencialmente

	rbio3.1	*db1*
/a/	96.25	3.75
/e/	97.5	2.5
/i/	97.5	2.5
/o/	100	0
/u/	100	0

Tabla 6.3: Porcentaje de registros para cada vocal

	db1	*bior1.5*	*rbio3.3*	*bior1.3*	*rbio1.1*
/a/	12.5	58.75	10	16.25	3.75
/e/	2.5	33.75	46.25	15	2.5
/i/	2.5	67.5	16.25	10	2.5
/o/	18.75	50	0	31.25	0
/u/	2.5	63.75	0	33.75	0

Tabla 6.4: Porcentaje de registros para cada vocal. Segundas mejores

es una de las más usadas en la literatura para tal fin.

6.2.2. Algoritmos de segmentación sonora/sorda

El algoritmo de segmentación sonora sorda basado en WT se compara con el que se encuentra en el *toolbox* de análisis de señales de voz de Childers [19], donde se usan ventanas de 250 muestras. En el algoritmo basado en WT se usan ventanas de la misma apertura temporal para facilitar la comparación, es decir de 9.1 ms. Es de notar que este algoritmo funciona bien con otros tamaños de ventana.

Para la separación de los sonidos sordos, se examina para cada banda la distribución de los coeficientes *wavelet*, obtenidos a partir de la FWT. Primero, se divide la señal de voz en 4 bandas diferentes, y la energía para cada banda es calculada en el dominio *wavelet*. El segmento de la señal de entrada es entonces clasificada como sonido sordo con el siguiente procedimiento [74]:

1. Use la FWT para encontrar W_3^A, W_3^D, W_2^D y W_1^D.
2. Calcule la energía promedio para cada subbanda EW_3^A, EW_3^D, EW_2^D y EW_1^D.
3. El segmento analizado es clasificado como sordo si se cumplen las siguientes condiciones:
   - $W_1^D > W_3^A$ y $W_1^D > W_3^D$ y $W_1^D > W_2^D$
   - $$\frac{EW_3^A}{EW_1^D} < 0{,}9$$

En el cálculo de la FWT, a modo de *wavelet* madre se usa la *wavelet* biortogonal *rbio3.1*, ya que resulta ser conveniente para la estimación del *pitch*.

La comparación de los resultados se realizó usando los criterios de falso positivo y falso negativo. Los falsos positivos corresponden a aquellas ventanas para las cuales se determina que son sonoras, sin serlo; y los falsos negativos corresponden a las ventanas que fueron etiquetadas como sordas sin serlo.

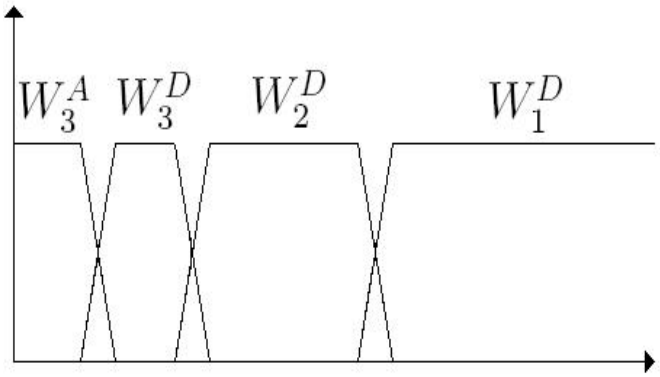

Figura 6.1: Niveles de descomposición usados para la segmentación sonora sorda

	WT	*Childers*
Falso positivo	0.2	2
Falso negativo	2.6	0.2

Tabla 6.5: Porcentaje de falsos positivos y falsos negativos para la segmentación sonora sorda usando DWT y Childers

En la prueba de los algoritmos se usaron registros de voz continua provenientes de la base de datos de LPH, tanto del sexo masculino como femenino. Los registros contienen palabras que agrupan los fonemas del tipo sonoro y sordo(africadas). Las señales están muestreadas a una frecuencia de $22050\ Hz$, a 16 bits. En las figuras (6.2) y (6.3) se observa el funcionamiento de los algoritmos en discusión. En la evaluación no se usaron segmentos de voz donde solo existiesen segmentos sonoros, en vez de ello se usan registros que además poseen porciones del tipo africadas brindando mayor rigidez a la confrontación de los algoritmos. Los resultados del anterior algoritmo se pueden observar en la tabla 6.5.
El algoritmo basado en TMA para la estimación del *pitch* puede ser usado para la segmentación de los tipos sonoro sordo, uno de sus resultados puede ser apreciado en la figura (6.2.2) . Usa el siguiente principio: dado que para el ruido los máximos locales de la WT estarán desordenadas las TMA serán de menor energía. Los resultados respecto a la segmentación, usando TMA, se aceptan. Se usaron 5 registros de 5 hablantes con 10 segmentos

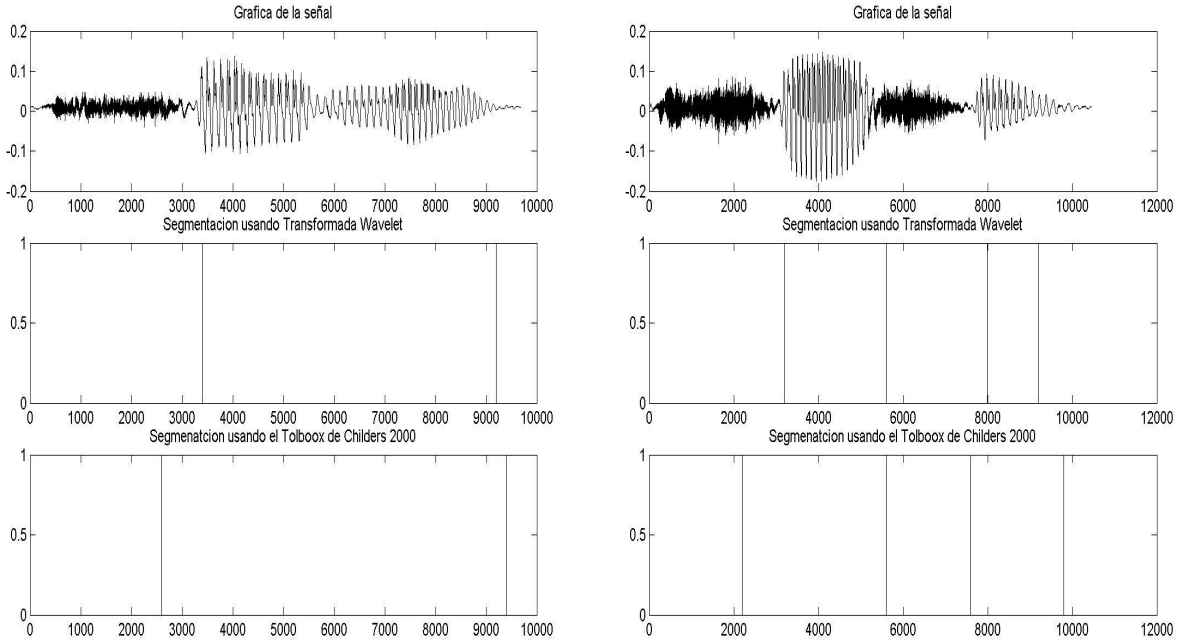

Figura 6.2: Segmentación sonora/sorda para la palabra /*cielo*/ y /*susi*/ respectivamente. Usando WT (arriba) y usando el *toolbox* de Childers (abajo)

	rbio3.1	db1
Falso positivo	22.5	25
Falso negativo	0	0

Tabla 6.6: Porcentaje de falsos positivos y falsos negativos para la segmentación basada en TMA

sonoros por cada hablante. Falso Positivo corresponde al número de segmentos en % que juzgó como sonoros, sin serlo, ver tabla 6.6. Pero éste algoritmo se descartó de la evaluación respecto a los otros dos algoritmos debido a su pobre funcionamiento en presencia de voces del tipo africadas.

El algoritmo basado en la DWT presenta mejor desempeño respecto a los falsos positivos, es decir, tiende a clasificar una menor cantidad de ventanas respecto a childers, pero entrega una mayor cantidad de ventanas clasificadas como falsos negativos, ver tabla (6.5), debido a que tiende a obviar aquellas ventanas que hacen parte del inicio y fin de palabra, sobre

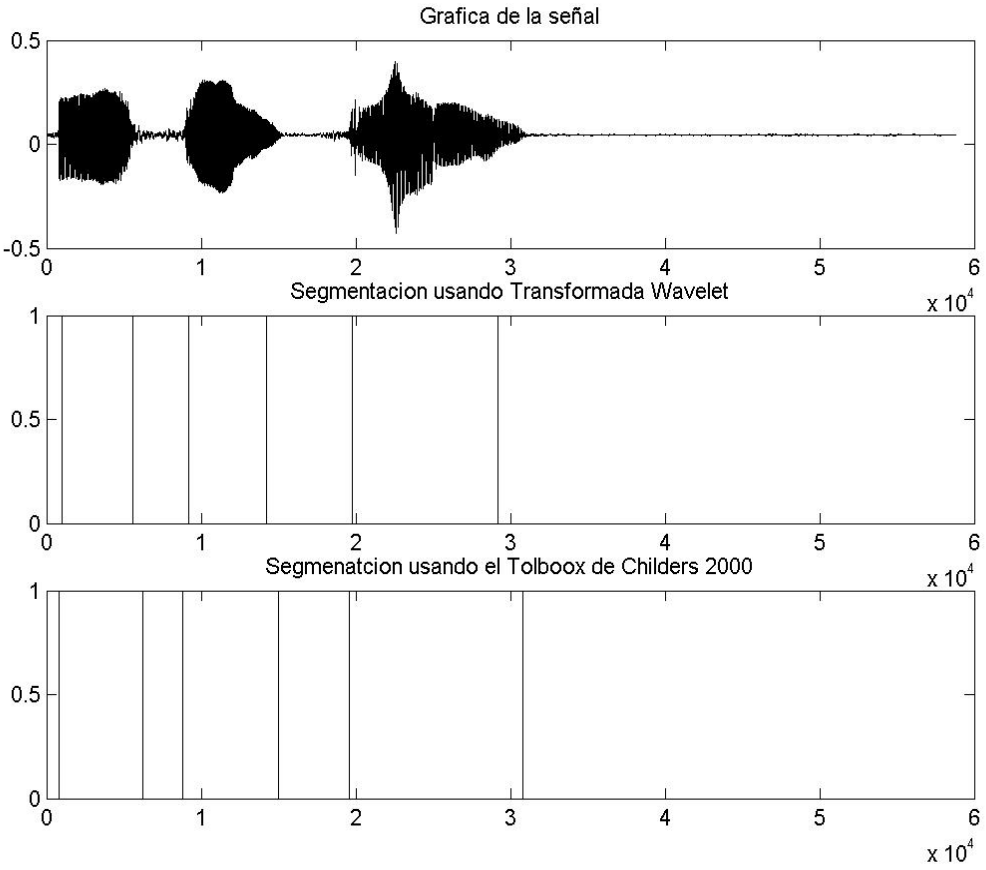

Figura 6.3: Segmentación sonora/sorda para un registro que contiene las palabras /*susi, cielo*/

todo si dichos instantes son de muy baja energía. Es de anotar además que el algoritmo basado en DWT es notoriamente menos costoso computacionalmente respecto de Childers.

6.2.3. Algoritmos de estimación del pitch

En estimación del pitch es suficiente tomar sólo unas cuantas escalas de la transformada *wavelet* continua, sin embargo, en el cálculo computacional, es preferible el empleo de algoritmos rápidos. En el presente trabajo se usó el algoritmo *Algoritme à Trous* o transformada *wavelet* diádica (DyWT, *Dyadic Wavelet Transform*) [55], el cual está basado en bancos de filtros. Para la estimación de la frecuencia fundamental se usaron dos algorit-

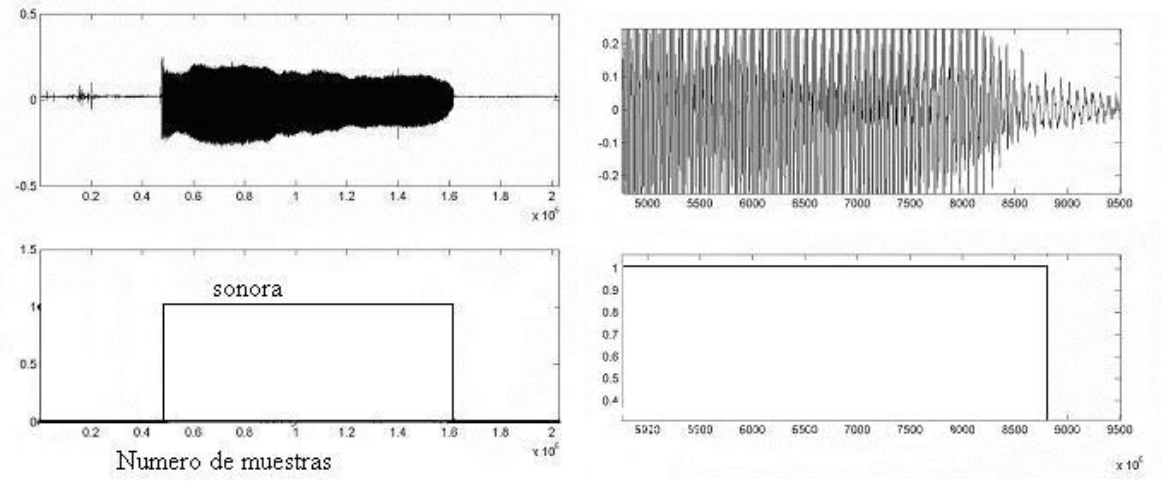

Figura 6.4: Vocal /a/. Segmentación basada en TMA. A la izquierda se aprecia una vista detallada

mos. El primero de ellos se basa en la formación de trayectorias de máxima amplitud y la segunda en la correlación de distancia entre máximos locales para diferentes escalas de descomposición.

Algoritmo basado en la determinación de trayectorias de máxima amplitud

1. *Determinación de los máximos locales*: Es la primera etapa del sistema y consiste en determinar el máximo valor para cada ventana de análisis. El tamaño de la ventana N_m se escoge asumiendo que la mayor frecuencia del Pitch que se pueda encontrar será de 500 Hz, lo cual nos da un mínimo periodo de pitch, con lo que se garantiza que para cada ventana existirá a lo sumo un máximo que corresponda a un GCI.

2. Determinación de las Trayectorias: Para ésta etapa se usa el algoritmo descrito por 5.1, cuya explicación se entregó anteriormente.

3. Verificación de la energía: Dentro de las trayectorias halladas, se descartan aquellas que sean menores respecto a sus vecinas, tal como se explica en 5.2.1.

4. Determinación de las distancias: A estas alturas solo resta determinar las distancias de las posiciones temporales de las TMA que corresponden a GCI.

En la figura 5.3 se muestra la estimación del pitch para la vocal */a/* perteneciente a un hablante femenino, que corresponde a la línea obscura. El método basado en TMA ofrece una buena estimación, pero debido a que require de un algoritmo basado en programación dinámica para la construcción de las TMA su desempeño es lento. Se comprueba que es sensible al ruido corroborando la afirmación dada en 5.2.1. Además se aprecia que la determinación de los GCI en su totalidad es complicado. Motivos por los cuales no se tendrá en cuenta en la evaluación.

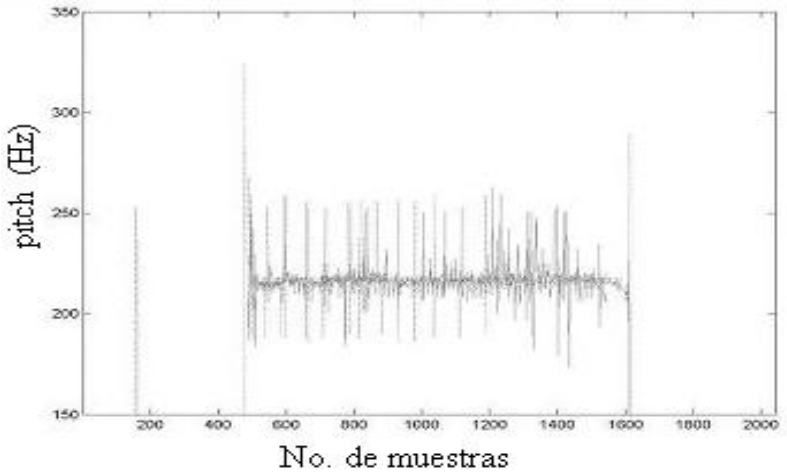

Figura 6.5: Fonema /a/ y su correspondiente pitch estimado. Línea obscura: Método de TMA. Línea a Tramos: HPS

Algoritmo basado en la correlación de distancias

El algoritmo propuesto en el presente trabajo, el cual está basado en la correlación de distancias, se puso especial empeño en la selección de los máximos locales dados por la DyWT ya que tal etapa es vital para dicho algoritmo. Los pasos que lo constituyen son los siguientes:

1. Para la selección de los máximos locales se sigue el mismo procedimiento de la primera etapa en el algoritmo anterior.

2. Corrección de los máximos: Se desarrolla en dos etapas:

a) Si se encuentra que dos máximos están separados por una distancia menor a N_m, entonces se descarta el menor de ellos.

b) En éste trabajo se encontró que generalmente entre dos máximos que corresponden a GCI se encuentran un tercer máximo, pero de menor altura respecto a la altura de sus vecinos. Particularmente se descartaron aquellos máximos cuya altura era menor que el $80\,\%$ del promedio de sus vecinos.

3. EL paso siguiente consiste en determinar las distancias entre los máximos para cada escala. Debido a que máximos locales erróneos no pueden eliminarse en un $100\,\%$, entonces se aplica un filtro mediana de tamaño 3 para eliminar aquellos valores anómalos del vector de distancias.

4. Para el presente sistema se tomaron 6 escalas de descomposición, de las cuales se toman aquellos 4 vectores de distancias que poseen menor desviación estándar respecto a su valor.

5. El resultado final se obtiene al promediar los valores de estimación del pitch entregados por cada escala.

6.2.4. Prueba de concordancia en la estimación pitch

Para decidir con objetividad sin dos estimaciones provienen de la misma función de densidad de probabilidad, se plantea la prueba de hipótesis. Este procedimiento objetivo debe basarse tanto en la información obtenida al investigar, como en el margen de riesgo que estamos dispuestos a aceptar si nuestro criterio con respecto a la hipótesis resulta incorrecto [76].

En esta sección se establece la hipótesis de que las salidas de los estimadores provienen de la misma distribución de probabilidad (H_0), para luego tratar de refutarla. Se comparan las funciones de distribución de probabilidad acumulativas para las poblaciones obtenidas a partir de la estimación del valor medio de la frecuencia fundamental usando Praat [63], el algoritmo de Childers y el método basado en *wavelets* propuesto en éste trabajo. Tales distribuciones se ilustran la figura 6.6. El algoritmo *Praat* correspon-

de al software de libre adquisición sustentado por el *Institute of Phonetic Sciences* de la universidad de Amsterdam. El software se encuentra disponible vía internet en la página *http://www.fon.hum.uva.nl/praat/*. En el presente capítulo nos referiremos al algoritmo de estimación del pitch desarrollado en [19] por los nombre de SIFT y Childers indistintamente.

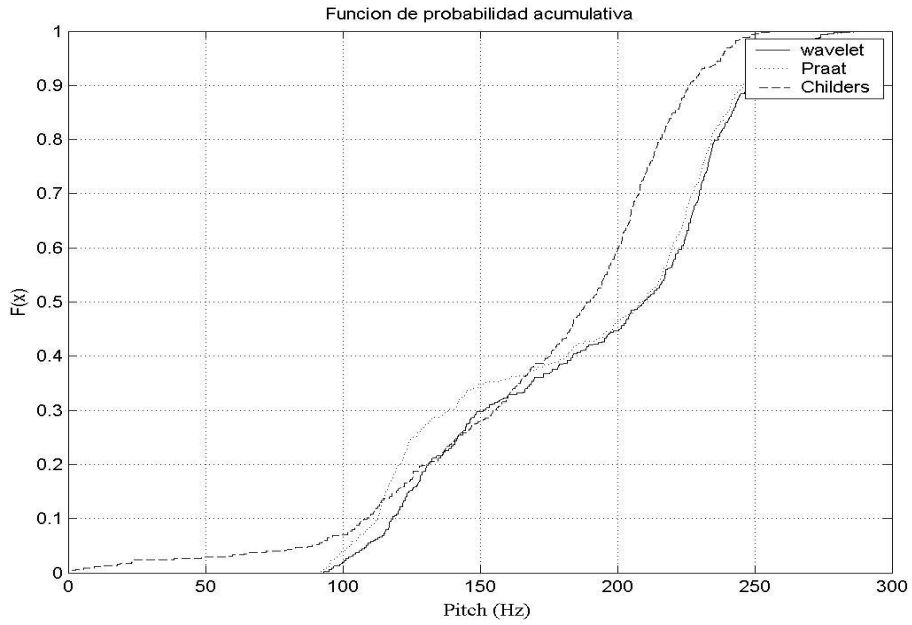

Figura 6.6: Funciones de distribución acumulativas para las tres estimaciones de la frecuencia fundamental: *wavelets, Praat y Childers*

Para la comparación se realiza la prueba de hipótesis usando el test de Kolmogorov-Smirnov, el cual es recomendado para datos continuos que son función de una sola variable [86], tomando de a pares hasta lograr comparar los métodos entre sí. Se escogen los niveles de significancia α más comúnmente usados $0{,}05$ y $0{,}01$ [76]. El resultado de tales comparaciones se muestra en la tabla 6.7 .

Al observar la gráfica (6.6) se aprecia el gran parecido en las estimaciones a partir de los 170 Hz entre nuestro método y el software *Praat*, y ya no tanto para las frecuencias menores a los 170 Hz. Además se aprecia en la tabla (6.7) que nuestro método pasa la prueba de hipótesis formulada mientras que el método expuesto en Childers no lo hace.

Poblaciones a comparar	Nivel de significancia α	Rechaza H_0
A con *B*	0.05	si
A con *B*	0.01	no
A con *C*	0.05	si
A con *C*	0.01	si
C con *B*	0.05	si
C con *B*	0.01	si

Tabla 6.7: Resultado de la prueba de hipótesis, confrontando todos contra todos. A: *wavelet*. B: *Praat*. C: Childers

6.2.5. Medida de similaridad respecto a la base de datos del pitch

En la tabla 6.8 se presentan las medidas de error cuadrático de los métodos SIFT de Childers y *wavelets* para cada hablante de la base de datos *Keele Pitch Database* respecto a la señal de referencia construida a partir de la señal del laringógrafo.

Para la obtener la medida de similaridad entre las estimaciones respecto a la base de datos de referencia del *pitch* se usaron señales tal y como se obtienen del proceso de registro, sin algún tipo de preprocesamiento. Al comparar nuestro algoritmo con el ofrecido en Childers y apreciando la tabla (6.8) se observa que nuestro algoritmo presenta mejor desempeño respecto a la medida de error cuadrático medio para las voces de la base de datos de referencia excepto para el hablante *f1*.

6.2.6. Pruebas de sensibilidad al ruido

Además de hicieron pruebas de ruido. A las señales se les agrega ruido blanco de distribución normal y se modifica su varianza de tal forma que se obtengan relaciones señal a ruido $SNR = 10 \log \frac{E\{s^2(t)\}}{E\{N^2(t)\}}$ [46] , de valores 0 dB, 6 dB, 10 dB y 20 dB. El operador E corresponde al valor esperado, $s(t)$ es la señal con ruido y $N(t)$ corresponde a la señal de ruido. En las tablas 6.9 y 6.10 se pueden apreciar los resultados para diferentes niveles de relación señal a ruido. Las figuras (6.7) y (6.8) corresponden a una traducción gráfica de las tablas

	sin ruido	sin ruido
	Childers	*wavelet*
f1 (1)	5313,6	8018,4
f2 (2)	978,6	243,9
f3 (3)	718,4	104,4
f4 (4)	1.136,7	132,5
f5 (5)	57,8	73
m1 (6)	10.100	1.180
m2 (7)	1.338	120,6
m3 (8)	3.350,4	63
m4 (9)	12.548	27,3
m5 (10)	6.365,6	33,3

Tabla 6.8: Medidas de error cuadrático medio los método expuesto en (Childers 2000) y correlación de distancias respecto a la referencia

anteriormente mencionadas. La medida de error se muestra en escala logarítmica para que sea posible su correcta visualización.

A partir de las figuras (6.7) y (6.8) se aprecia el buen desempeño del algoritmo basado en *wavelets* respecto del algoritmo presentado en [19] ante condiciones de ruido aditivo. Se observa que el error cuadrático medio de la estimación usando *wavelets* es menor para la mayoría de los hablantes (2,3,4,6,7,8,9,10), es decir para f2, f3,f4, m1, m2, m3, m4 y m5, y no lo es para los hablantes 1 y 5 (f1 y f5).

6.3. Parámetros de ruido

De una u otra forma los parámetros acústicos de ruido lo que hacen es medir la cantidad de componente de armónico que posee una señal de voz. Para extraer la parte coherente de la señal de voz se usa un método basado en la transformada WP y su capacidad para representar eficientemente las señales de voz.

	0 dB	0dB	6 dB	6dB
	Childers	Wavelet	Childers	wavelet
f1 (1)	2301800	3439700	76926.0	839760.0
f2 (2)	4067.2	1588.0	1670.8	472.9
f3 (3)	4496.3	645.4	1361.0	353.5
f4 (4)	6803.6	2286.6	2870.6	1070.5
f5 (5)	610.4	1137.0	101.4	353.3
m1 (6)	19521.0	2265.1	15925.0	2213.7
m2 (7)	8308.0	553.1	8749.7	202.4
m3 (8)	10059.0	1226.2	5499.2	228.4
m4 (9)	12934.0	167.43	10577.0	96.4
m5 (10)	13964.0	705.9	10558.0	180.9

Tabla 6.9: Medidas de error cuadrático medio los métodos SIFT y Wavelets respecto a la referencia

	10 dB	10dB	20 dB	20 dB
	Childers	Wavelet	Childers	wavelet
f1 (1)	2435.8	3154.9	1225.5	1322.0
f2 (2)	1851.4	288.7	1002.0	192.3
f3 (3)	881.0	186.2	725.8	95.6
f4 (4)	2606.8	488.3	1522	322.8
f5 (5)	94.5	108.0	85.7	91.2
m1 (6)	14944.0	1730.6	8025.8	1600.1
m2 (7)	3353.6	87.0	2202	80.5
m3 (8)	1466.0	51.7	1320.4	47.0
m4 (9)	10921.0	68.3	9253.0	51.7
m5 (10)	8693.2	108.7	5354.3	95.8

Tabla 6.10: Medidas de error cuadrático medio los métodos SIFT y Wavelets respecto a la referencia

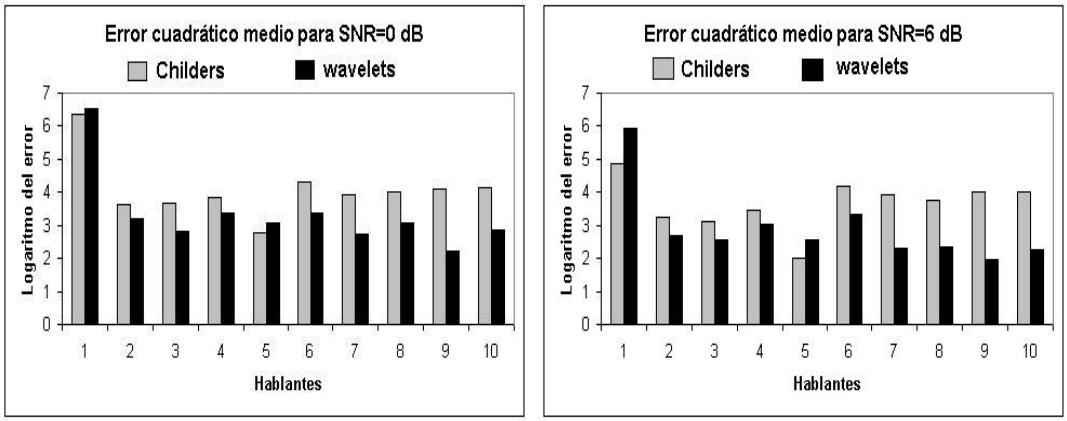

Figura 6.7: Medidas de ruido en escala logarítmica

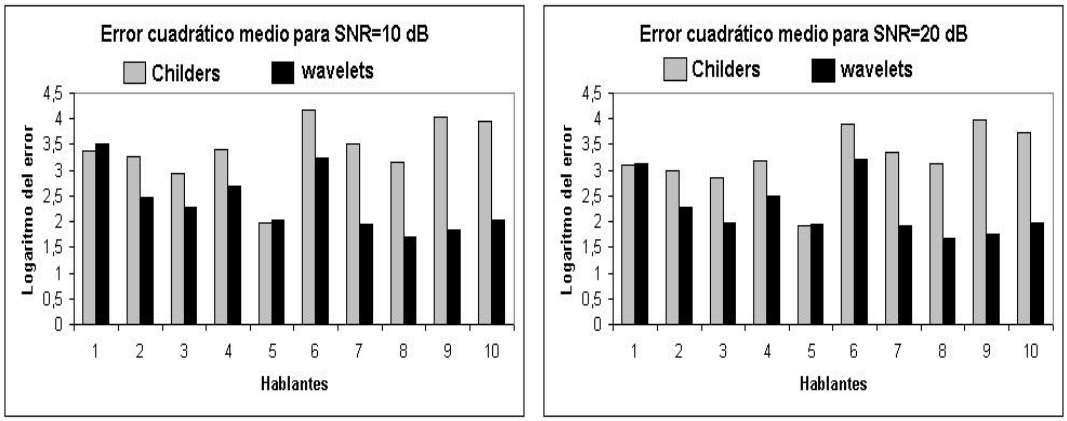

Figura 6.8: Medidas de ruido en escala logarítmica

El algoritmo mostrado en [88], [1] y [22] se explica a continuación, en el que se tiene en cuenta que los parámetros β y δ pueden ser ajustados:

1. Dada una señal de longitud N, se aplica el algoritmo de *Best Basis Selection* para obtener entre ellas la que mejor represente la señal desde el punto de vista de una función de coste de información

2. Luego se ordenan los coeficientes de esta base en forma decreciente.

3. El paso que sigue es establecer la cantidad de coeficientes que son significantes para la representación. La primera forma consiste en establecer aquel punto en la cual la derivada de los coeficientes ordenados presenta mayor cambio. La segunda es

más simple y consiste en tomar aquellos cuya magnitud sea mayor a determinado porcentaje de la mayor magnitud de los coeficientes.

4. Para la señal a analizar se halla la dimensión teórica, que es un tipo de función de coste de información. Se dice que la señal que se está analizando es incoherente si su dimensión teórica es menor que determinado umbral β. Esta medida sirve para rechazar una mala compresión, aún después de haber aplicado el algoritmo de *Best Bases*. Esta condición determina la condición de parada del algoritmo. Para este caso en particular se usa el mismo valor de la entropía.

5. Si la señal es no incoherente, es decir si aún posee componentes que pueden ser reconocidos como señal, entonces se descompone f en $c_1 + r_1$, donde c_1 es reconstruido a partir de los δN, $0 < \delta \leq 1$, coeficientes más grandes, mientras que r_1 es el residuo reconstruido a partir de los restantes coeficientes más pequeños. El proceso es iterativo, y la señal residual retorna y alimenta el paso 1 del algoritmo.

En las figuras (6.9) (6.10) (6.11) y (6.12) se muestra la energía coherente por registro, para voces adultas distribuidas en dos clases, la primera clase corresponde a las muestras de la 1 a la 40 (voces normales), y las restantes 51 muestras pertenecen a la segunda clase (voces disfónicas). Se observa que el valor de la energía coherente para cada una de las clases es significativamente diferente, por lo cual se deuce que podría fácilmente ser utilizada para la determinación de la disfonía.

6.4. Análisis discriminante aplicado a las características de representación

Para la extracción de características se usan dos tipos de transformadas *wavelet* la DyWT y la transformada *Wavelet Packet*. Las características son las siguientes:

– Coeficientes de la Transformada Wavelet Diádica (DyWT)

  La transformada *wavelet* diádica mantiene la invarianza a la traslación muestreando únicamente el parámetro escala de la CWT. En muchos métodos existentes, los val-

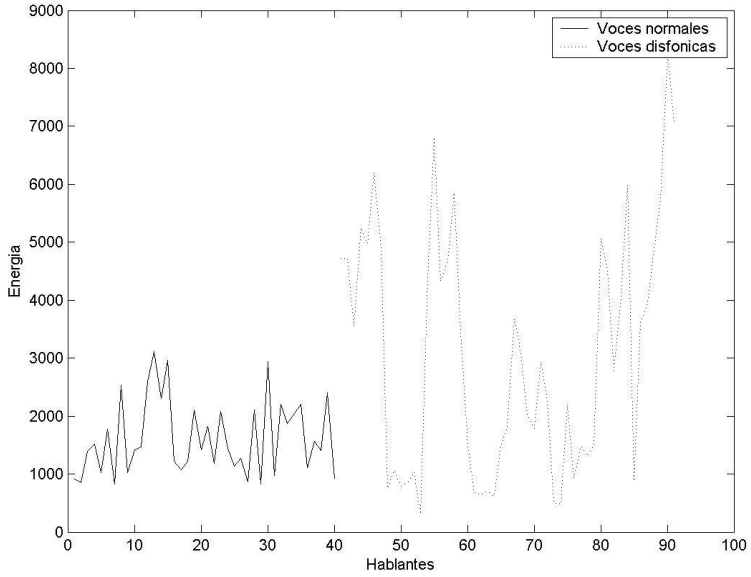

Figura 6.9: Energía coherente para la vocal /*a*/

ores de energía o los coeficientes *Wavelet* en sí mismos para todas las escalas o para las escogidas son tomados como características esenciales de la señal.En el presente trabajo se toman los coeficientes de mayor magnitud a manera de características; por cada escala se toman 2, cuidando de que estén alejados al menos por un periodo del pitch entre sí, y así evitar tomar dos coeficientes correspondientes a una misma singularidad; para un total de 12 características dado que se analizan 6 escalas.

– Valores de entropía de los subespacios de la transformada WP

  Para la prueba se tomaron aquellos segmentos estables para cada vocal sostenida que corresponde 2048 muestras. Al igual que en [7] y [6], se usan 5 niveles de descomposición usando la *Symlet* de orden 5 y la entropía de Shannon a modo de función de costo. Los valores de entropía resultantes, en total 63, vienen a conformar el espacio de características.

– Coeficientes de las bases discriminantes

  Los coeficientes de las bases discriminantes se obtienen al aplicar el algoritmos de

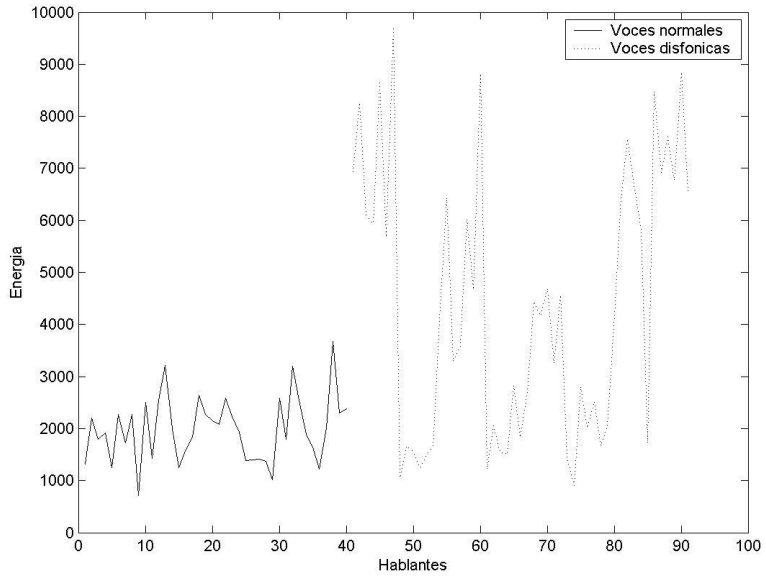

Figura 6.10: Energía coherente par la vocal /*e*/

Local Discriminant Bases descrito anteriormente. La señales que alimentan el algoritmo de *Local Discriminant Bases* corresponden a los segmentos estables de las vocales sostenidas, en particular, se toman las 2048 muestras de la parte más estable.

Similar al algoritmo de *best basis*, sea B_j^p el conjunto de vectores base en el subespacio W_j^p. Además, sea A_j^p la LDB restringida al *span* de B_j^p. Y, sea Δ_j^p una variable que contiene la medida discriminante del subespacio W_j^p. El Algoritmo LDB se expone a continuación:

1. Escoja un diccionario de bases ortonormales $\mathcal{D}$ y especifique el máximo nivel de descomposición J y medida discriminante aditiva D .
2. Construya los mapas de energía de tiempo-frecuencia Γ_c para las clases $c = 1, ..., C$
3. Haga $A_j^p = B_j^p$
4. Determine el mejor subespacio A_j^p para $j = J - 1, ..., 0$, $k = 0, ..., 2^j - 1$, de la siguiente forma:

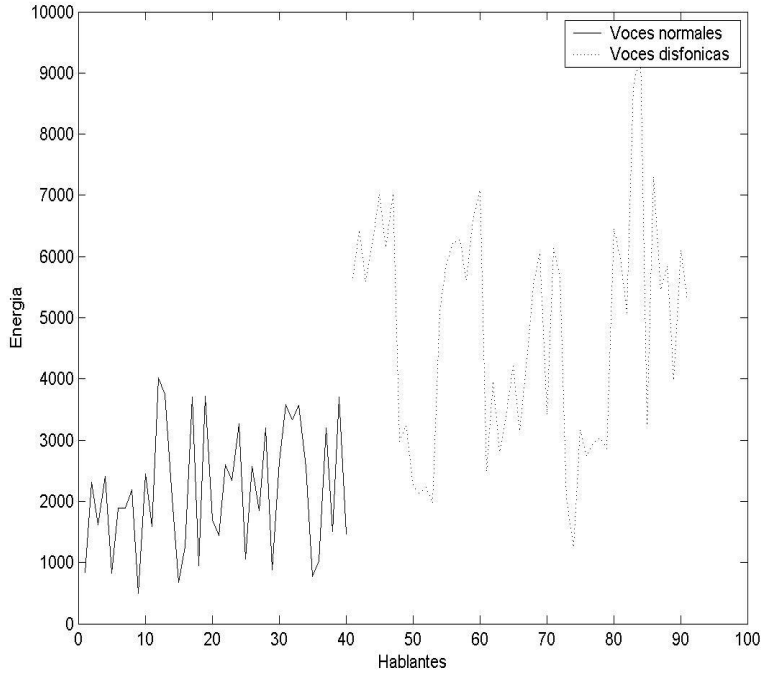

Figura 6.11: Energía coherente par la vocal /*i*/

si $\Delta_j^p \geq \Delta_{j+1}^{2k} + \Delta_{j+1}^{2k+1}$ entonces $A_j^p = B_j^p$, de lo contrario $A_j^p = A_{j+1}^{2p} \oplus A_{j+1}^{2p+1}$ y haga $\Delta_j^p = \Delta_{j+1}^{2k} + \Delta_{j+1}^{2k+1}$

5. Ordenar las bases respecto a su poder discriminante
6. Usar los $k(\leq n)$ bases de mayor poder discriminante para alimentar los clasificadores

Las bases obtenidas mediante le algoritmo de LDB se muestran en las figuras (6.13), (6.14), (6.16), (6.4) y (6.17), para las vocales /*a*/,/*e*/,/*i*/,/*o*/ y /*u*/ respectivamente. Para disminuir el espacio de características, se tomaron solo aquellas bases para las cuales su divergencia resultó mayor. Adicionalmente se aplicó la medida de discriminancia lineal a cada coeficiente para luego escoger los de mayor discriminancia, estos últimos son los que alimentan el clasificador, es ésta la estrategia que Saito (autor de este estrategia, en compañía de R. R. Coifman) usa originalmente en sus trabajos.

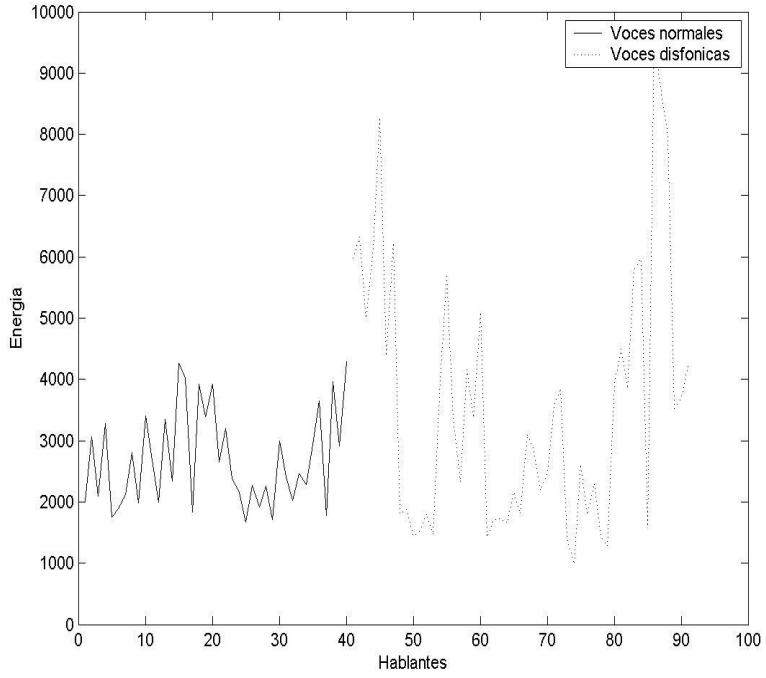

Figura 6.12: Energía coherente par la vocal /*o*/

Análisis discriminante . Se presentan los diagramas de dispersión de clases para dos coeficientes obtenidos a partir de los valores de entropía de la transformada WP en la figura en la figura (6.18), en la figura (6.19) se presente lo correspondiente al algoritmo de LDB y en la figura (6.20) las características derivadas de los máximos de los espacios discriminantes obtenidos al aplicar LDB.

Para evaluar el desempeño de las características se aplicó análisis discriminante lineal, y el clasificador para esas características fue evaluado con validación cruzada de $n = 1$. Los resultados se muestran en las tablas (6.11) y (6.12). Para obtener la primera columna de la tabla (6.12) lo que se hace es tomar la máxima magnitud de cada subespacio discriminante, mostrados en las figuras (6.13), (6.14), (6.16), (6.4) y (6.17). Para obtener la segunda columna se toman dos máximos y para la tercera se toman también dos máximos, pero cuidando de que cada uno corresponda a cada una de las mitades del vector que describe el subespacio discriminante. Los resultados se aceptan, partiendo del hecho de que el clasificar que se usa es lineal, y de hecho se esperan mejores resultados al usar esquemas

	DyWT		*Entropías WP*		*LDB original*	
vocal	% clasificación	No. par	% clasificación	No. par	% clasificación	No. par
/*a*/	78,9	12	77,8	63	70,3	78
/*e*/	81,1	12	67,8	63	61,5	86
/*i*/	80	12	72,2	63	69,2	34
/*o*/	85,5	12	68,9	63	50,5	85
/*u*/	75,5	12	61,1	63	58,2	35

Tabla 6.11: Porcentaje de clasificación para voces disfónicas usando DyWT, los valores de entropías la transformada WP y LDB en su esquema original

	LDB (un máximo)		*LDB (dos máximos)*		*LDB (máximos separados)*	
vocal	% clasificación	No. par	% clasificación	No. par	% clasificación	No. par
/*a*/	78	5	79,1	10	76,9	10
/*e*/	76,9	8	78	16	75,8	16
/*i*/	76,9	8	74,7	16	76,9	16
/*o*/	60,4	1	60,4	2	63,7	2
/*u*/	56	1	61,5	2	64,8	2

Tabla 6.12: Porcentaje de clasificación para voces disfónicas usando LDB (un máximo), LDB (dos máxomos) y LDB (dos máximos separados)

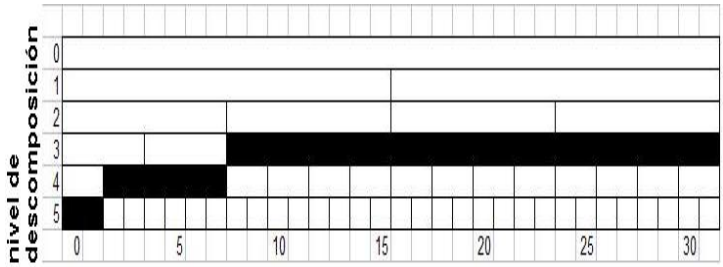

Figura 6.13: Subespacios seleccionados como LDB para la vocal /a/

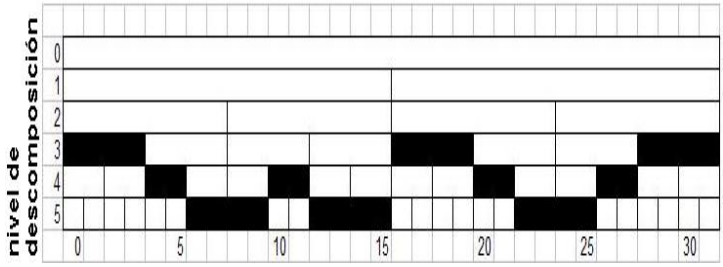

Figura 6.14: Subespacios seleccionados como LDB para la vocal /e/

de clasificación más elaborados.

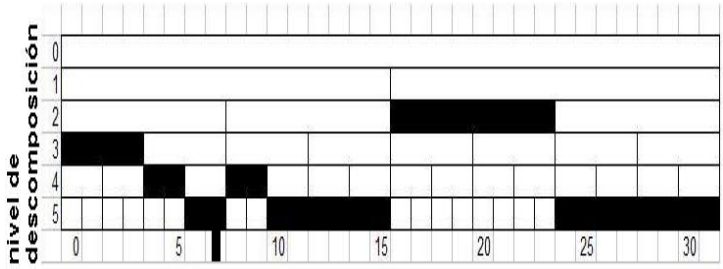

Figura 6.15: Subespacios seleccionados como LDB para la vocal /i/

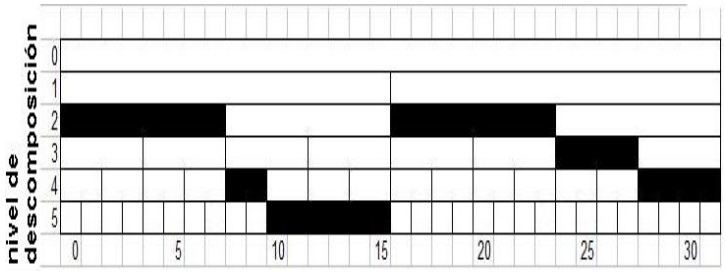

Figura 6.16: Subespacios seleccionados como LDB para la vocal /o/

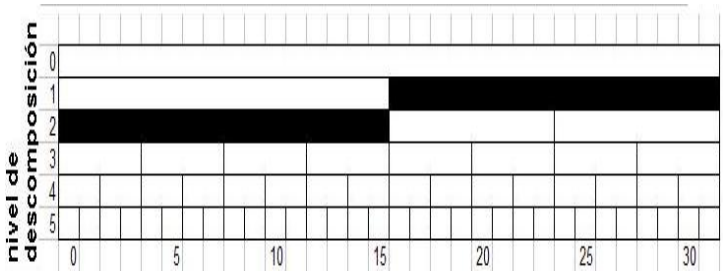

Figura 6.17: Subespacios seleccionados como LDB para la vocal /u/

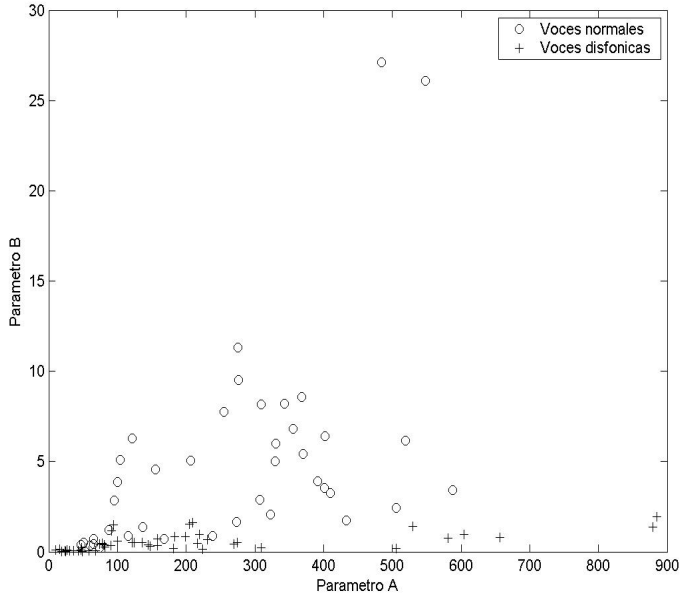

Figura 6.18: Parámetros basados en entropías de la transformada WP

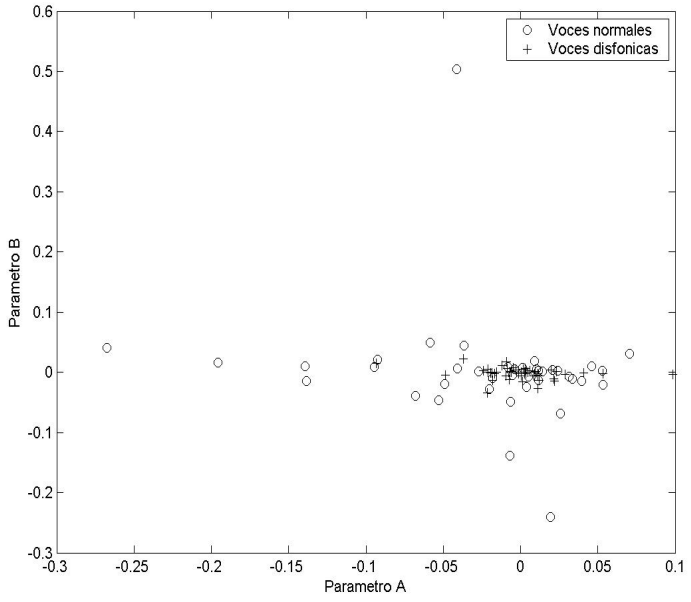

Figura 6.19: Parámetros basados en los coeficientes obtenidos a partir de Local Discriminant Bases

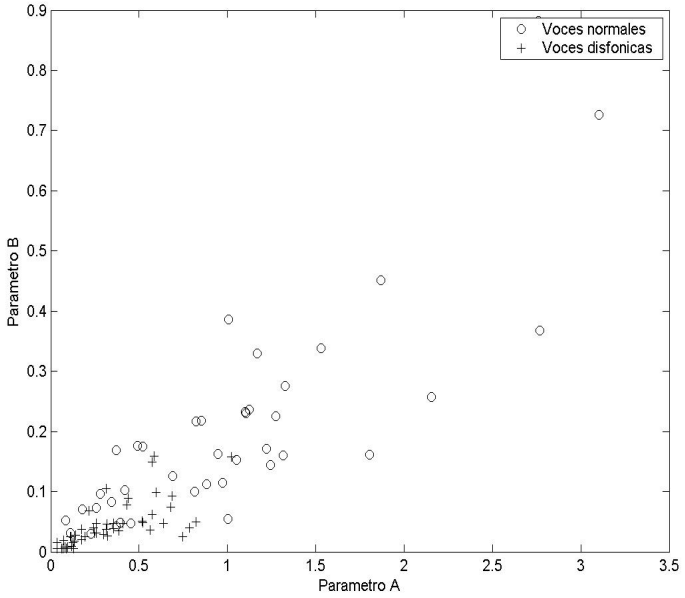

Figura 6.20: Tomando máximos de los subespacios discriminantes

Capítulo 7

Conclusiones

En primer lugar, se propone un nuevo método de estimación de la frecuencia fundamental de señales de voz El método fue probado respecto de la base de datos *Keele Pitch Database*, logrando resultados aceptables, tal como se aprecia en la tabla 6.8. Su comportamiento ante esta base de datos fue comparado con una variante del método SIFT desarrollado en [19] logrando obtener mejores resultados respecto a éste último método. Aparte de la base de datos del *pitch Keele* y de la variante del método SIFT, existió un tercer juez cuya finalidad era corroborar los resultados, el *Praat*. Las funciones de distribución mostraron un comportamiento notoriamente similar, ver figura (6.6), además de pasar la prueba de hipótesis, ver tabla 6.7 . El algoritmo desarrollado puede trabajar aún en ambientes de ruido, sin necesidad de aplicar algún tipo de filtración ó de-noising, ver tablas (6.9, 6.10). Además el tiempo de proceso respecto al método es notoriamente menor.

En segundo lugar, para la selección de la wavelet madre se desarrolló una metodología basada en la combinación de elementos teóricos y empíricos en la selección de cual wavelet podría entregar mejor rendimiento para la estimación de la frecuencia fundamental. Los elementos teóricos colaboran en reducir la cantidad de posibilidades de wavelet. Una vez en éste nivel se procede a aplicar una medida de costo tendiente a disminuir la cantidad de máximos locales en las escalas de descomposición de la WT que no corresponden a GCI. Los resultados mostrados en la tabla (6.3) reportan a la wavelet de la familia spline $rbio3{,}1$ como la más opcionada en la estimación del pitch, que coincide con ser una de las más usadas en la literatura para la estimación del pitch.

En tercer lugar, se discute la viabilidad de clasificar voces de tipo anormal(voces disfónicas), usando características basadas en coeficientes *wavelet*. Se usan varios esquemas para la toma de los coeficientes *wavelet*, entre ellos la DyWT y el establecimiento de subespacios discriminantes hallados a partir del algoritmo de *Local Discriminant Bases*, entregando buenos resultados en la clasificación de voces patológicas. Ver tablas (6.11) y (6.12).
En cuarto lugar, se desarrollan métodos de segmentación de los tipos sonoro y sordo usando WT. El método propuesto en [74] es notoriamente superior en cuanto tiempo de cómputo y confiabilidad respecto del método basado en TMA, y comparable en desempeño con el ofrecido en [19], ver tabla 6.5.

En quinto lugar, debido a la pobre definición de lo que es ruido en las señales de voz se propone extraer la estructura coherente de las señales de voz a modo de parte armónica, basado en el algoritmo de *best bases selection*. Su eficiencia para la clasificación de voces disfónicas se prueba, entregando buenos resultados preliminares desde el punto de vista discriminante, ver figuras (6.9), (6.10), (6.11) y (6.12) .

Bibliografía

[1] *Wavelet feature extraction for discriminant tasks*.

[2] *Estimation by means of wavelet analysis of the signal-to-noise ratio of dysphonic voices*, 1997.

[3] C. G. A. Croisier, D. Esteban. Perfect channel splitting by use of interpolation, decimation, tree decomposition techniques. *Int. Conf. on Information Sciences/Systems*, Aug. 1976.

[4] D. M. R. A. Mojsilovic, M. V. Popovic. On the selection of an optimal wavelet basis for texture characterization. *IEEE Transactions on Image Processing*, 9(12), December 2000.

[5] W. A. P. A. Said. An image multiresolution representation for lossless and lossy compression. *IEEE Trans. on Image Pro.*, pages 1303–1010, 1996.

[6] V. G. A. Schuck and J. Wisbeck. Dysphonic voice classification using wavelet packet transform and artificial neural network. *The 25th annual internationational conference of the IEEE Engineering in Medicine and Biology Society, México 2003*, 2003.

[7] A. P. A. Shuck Jr. On the use of wavelet packet transform as a feature extractor for pathological voice assessment. In *IFMBE(IMEBC02)*, volume 3, pages 506–507, December 2002.

[8] P. S. Addison. *The illustrated Wavelet TRansform Handbook*. Institute of Physics Publishing, 2002.

[9] R. Alzate. Estimación del pitch en tiempo real orientado al aav. Trabajo de grado, Universidad Nacional de Colombia, Manizales, 2003.

[10] B. W. G. an Les E. Atlas. Optimizing time-frequency kernels for classification. *IEEE. Trans. on Signal Processing*, 49(3), 2001.

[11] V. D. B. Yegnanarayana, C. d'Alessandro. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Trans. on speech and auudio proc.*, 6(1), 1998.

[12] P. D. Beng T. Tan, Minyue Fu. The use of wavelet transforms in phoneme recognition. Technical report, University of Newcastle and National Acoustic Laboratories, 1996.

[13] Boies. *Otorrinolaringología de Boies: Enfermedades del oído, vías nasales y laringe*. Nueva Editorial Interamericana, 1981.

[14] V. S. Boualem Boashash. *Waveltes and Signal Processing*, chapter High Performance Time-Frequency Distributions for Practical Applications. BirkHauser, 2003.

[15] P. Brémaud. *Mathematical Principles of Signal Processing: Fourier and Waveelt Analysis*. Springer, 2001.

[16] A. Bultheel. *Wavelets with applications in signal- and image processing*. http://www.cs.kuleuven.ac.be/ ade/WWW/WAVE/contents.html, 2001.

[17] R. A. G. C. Sydney Burrus and H. Guo. *Introduction to Wavalets and Wavelet Transform*. Prentice Hall, 1998.

[18] R. Chandramouli and K. Ramachandran. *Wavelets and Signal Processing*, chapter Wavelts for Statistical Estimation and Detection, pages 106–107. BirkHauser Boston, 2003.

[19] D. G. Childers. *Speech Processing and Synthesis Toolboxes*. John Wiley and Sons, 2000.

[20] L. Cohen. *Wavelets and Signal Processing*, chapter The Wavelet Transform and Time-Frecuency Analysis. BirkHauser, 2003.

[21] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Info. Theory*, 38(2), 1992.

[22] R. R. Coifman and M. V. Wickerhauser. Experiments with adapted wavelet de-noising for medical signals and images. In *Time-Frequency and Wavelets in Biomedical Engineering*. IEEE Press, 1998.

[23] M. F. D. Michaelis and H. W. Strube. Selection and combination of acooustic features for the description of pathologic voices. Technical report, Drittes Phisykalisches Institut, 1998.

[24] I. Daubechies. *Ten Lectures on Wavelets*. Notes from the 1990 CBMS-NSF Conference on Wavelets and Applications at Lowell, MA, 1992.

[25] L. Debnath. *Wavelet Transforms and their Applications*. BirkHauser, 2002.

[26] A. A. Fracois Le Huche. *Patología Vocal: Semiología y disfonías disfuncionales*, volume 2 of *LA VOZ*. MASON, 1994.

[27] S. Furui. *Digital Speech Processing, Synthesis and Recognition*. Mercel Dekker, Inc., 1989.

[28] L. J. García. *Transformada wavelet aplicada a la extracción de información ennseñales de voz*. PhD thesis, Univesitat Politecnica de Catalunya, 1998.

[29] B. Gold and N. Morgan. *Speech and Audio Signal Processing*. John Wiley and Sons, 2000.

[30] A. Graps. An introduction to wavelets. *IEEE Computational Science and Engineering*, 2(2), 1995.

[31] J. Hansen. A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment. *IEEE transactions on Biomedical Engineering*, 8(2):106–122, 1994.

[32] C. Herley. *Digital Signal Processing Handbook*, chapter Wavelets and Filter Banks. Chapman and Hall/ CRCnetBASE, 1999.

[33] H. H.-T. W. Hsi-Tsung. "a glottal excited linear prediction (gelp) model for low-bit-rate speech coding". *National Taiwan University of Science and Technology. Taipei, TAIWAN*.

[34] K. F.-J. H. H.-T. W. Hsin-Jen. "a pseudo glottal excitation model for the linear prediction vocoder with speech signals coded at 1.6 kbps". *IEICE Transactions. N° 8, Vol E83-D*, Agosto de 2000.

[35] M. A. J. Donnell, R. Rohrich. Velopharyngeal incompetence:: A guide for clinical evaluation. *Plastic and reconstructive surgery*, Dec. 2003.

[36] j. L. M. J. Gonzáles, T. Cervera. Análise acústica da voz captada na faringe próximo á fonte glótica através da microfone acoplado ao fibrolaringoscópido. *Revista Brasileira de Otorrinolaringología*, 67, 2001.

[37] j. L. M. J. Gonzáles, T. Cervera. Análisis acústico de la voz: Fiabilidad de un conjunto de parámetros multidimensionales. *Acta Otorrinolaringológica española*, (53):256–268, 2002.

[38] Y. S. J K. Shah, N. Iyer and E. Yantorno. Robust voiced/unvoiced classification using novel features and gaussian mixture model. Technical report, Speech Processing Lab / ECE Dept. / Temple University, 2003.

[39] P. P.-P. J. Polygiannakis and X. Moussas. On signal-noise decomposition of time-series using the continuous wavelet transform: application to sunspot index. *Monthly Notices of the Royal Astronomical Society*, 343:725–734, 2003.

[40] G. S. J. Stegmann and K. A. Fischer. Robust classification of speech based on the dyadic wavelet transform with application to celp coding. In *ICASSP 96*, pages 546–549, 1996.

[41] L. Janer. Modulated gaussian wavelet transform based speech analyser(mgwtsa) pitch detection algorithm (pda). In *EUROSPEECH*, 1995.

[42] M. Jansen. *Noise Reduction by Wavelet Thresholding*. Springer-Verlag, 2001.

[43] P. T. Kadambe, S. Srinivasan and H. Szu. Representation and classification of unvoiced sounds using adaptive wavelets. Technical report, AT T Bell Laboratories, 1993.

[44] S. Kadambe and G. F. Boundreaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *IEEE Trans. on Info. Theory*, 38(2), 1992.

[45] S. Kadambe and G. Bourdeaux-Bartels. A comparison of a wavelet functions for pitch detection of speech signals. *International Conference on Acoustics, Speech, and Signal Processing*, 1991.

[46] A. B. K.S. Shanmugan. *Random signals: detection, estimation and data analysis*. Jhon Wiley and Sons., 1988.

[47] F. le Huche. *La Voz.* Patología Vocal: Semiología y disfonías disfuncionales, 1994.

[48] P. Lee. Wavelet filter banks in perceptual audio coding. Master's thesis, University of Waterloo, Ontario, Canadá, 2003.

[49] P. Lieberman. Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. *J. Acoust. Soc. Am.*, 1963.

[50] I.Ñ. M. Baseville. *Detection of Abrupt Changes: Theory and Applications*. Prentice-Hall Inc., 1993.

[51] H. S. M. Frohlich, D. Michaelis. Acoustic breathiness measures in the description of pathologic voices. Technical report, Universitat Gottingen, 1999.

[52] T. S. M. Sakamoto. An automatic pitch-marking method using wavelet transform. In *ICSLP2000*, 2000.

[53] P. T. M. Unser and A. Aldroubi. Shift-orthogonal wavelet basis. *IEEE Trans. on Sig. Proc.*, 46(7), July 1998.

[54] I. R. A. MacKay. *Phonetics: The Science of Speech Production*. Allyn and Bacon, 1987.

[55] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[56] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. on Patt. Anal. and Mach. Int.*, 1989.

[57] F. P. G. Meyer and W. Ainsworth. A pitch extraction reference database. In *Eurospeech95*, http://www.liv.ac.uk/Psychology/HMP/projects/pitch.html, 1995.

[58] D. D. N. Gonzáles. Application of singularity detection with wavelets for pitch estimation of speech signals. In *EUROSPEECH94*, 1994.

[59] F. G. F. W. N. Saito, R. Coiffman. Discriminant feature extraction using empirical probability density estimation and a local basis library. *Pattern Recognition*, 2002.

[60] D. E. Newland. Practical signal analysis: do wavelets make any difference? In *Design Engineering Technical Conferences,DETC'97*, 1997.

[61] V.Ñgoc and C. d'Alessandro. Robust glottal closure instant detection using the wavelet transform. *LIMSI-CNRS, Orsay, France*, 1999.

[62] F. Ojeda. Extracción de características usando transformada wavelet en la identificación de voces patológicas. Trabajo de grado, Universidad Nacional de Colombia, Manizales, 2003.

[63] F. Ojeda. Praat: doing phonetics by computer. http://praat.org, 2004.

[64] A. Parraga. *APLICAÇÃO da Transformada Wavelet Packetna Análise e CLASSIFICAÇÃO de sinais de vozes patológicas*. PhD thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

[65] S. Pittner and S. V. Kamathi. Feature extraction from wavelet coefficients for pattern recognition tasks. *IEEE Trans. on Patt. Analysis and Mach. Intell.*, 21(1), 1999.

[66] R. L. Queiroz. *Digital Signal Processing Handbook*, chapter Lapped Transforms. Chapman and Hall/ CRCnetBASE, 1999.

[67] J. L. F. R. E. Croisier, S. A. Weber. Digital coding of speech in subbands. *Bell System Technical J.*, Oct. 1976.

[68] N. Saito. *Local Feature Extraction and its Applications Using a Library of Basis*. PhD thesis, Yale University, 1994.

[69] N. Saito. Local discriminant bases and their applications. *Mathematical Imaging and Vision*, 5(4):337–358, 1995.

[70] N. Saito. Classification of acoustical geofisical waveforms using time-frequency atoms. In *Computing Section of Amer. Statist. Assoc.*, pages 322–327, 1997.

[71] N. Saito and R. Coifman. Local discriminant bases. In *Wavelet Applications in Signal and Image Processing, SPIE 2303*, pages 2–14, July 1994.

[72] N. Saito and R. Coifman. Extraction of geological information from acoustic well-logging waveforms using time-frequency wavelets. *Geophysics*, 62(6), 1997.

[73] N. se. *Juan sin Nombre*. No se, No se.

[74] J. W. Seok and K. S. Bae. Speech enhancement with reduction of noise components in the wavelet domain. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-97*, 2:1323–1326, 1997.

[75] J. M. Shapiro. Embebeded image coding using zerotrees of wavelet coefficients. *IEEE Tans. on Signal Proc.*, Dec. 1993.

[76] A. Sigarroa. *Biometría y diseño experimental*. Editorial pueblo y educación, 1985.

[77] M. M. Sondhi and J. Shroeter. *Digital Signal Processing Handbook*, chapter Speech Production Models and their Digital Implementations. Chapman and Hall/ CRCnet-BASE, 1999.

[78] Y. M. Stéphane Jaffard and R. D. Ryan. *Wavelets: Tools for Science and Technology*. Society for Industrial and Applied Mathematics, 2001.

[79] R. V. Subhasis Saha. Adaptive wavelt filter in image coders: How important are they. Technical report, University of California, Davis, 1999.

[80] X. Sun. *(—RARO—)A Pitch Determination Algorithm Based on Subarmonic to Harmonic Ratio*. 2000.

[81] N. V. Thakor and D. Sherman. *Biomedical Signal Analysis*, chapter Wavelet (Time -Scale) Analysis in Biomedical Signal Processing. CRC Pres. Inc., 1995.

[82] B. Torrésani. An overview of wavelet analysis and time-frecuency analysis. Technical report, LATP, CMI, Université de Provence, 1898.

[83] F. Vargas. Selección de características en el análsis acústico de voces. Master's thesis, Universidad Nacional de Colombia, Manizales, 2003.

[84] B. Vidacovic and P. Muller. *Wavelets for kids: A tutorial introduction*. Duke Univesity.

[85] O. A. V. S. R. W. *Digital Signal Processing*. Prentice Hall, p. 480 - 531, New Jersey. USA, 1975.

[86] S. T. W. Press, B. Flannery and T. Vetterling. *Numerical Recipes in C: The Art of Science Computing*. Cambridge University Press, p. 564 - 572, Cambridge. UK, 1993.

[87] C. Wendt and A. P. Petropulu. Pitch determination and speech segmentation using the discrete wavelet transform. *IEEE International Symposium on Circuits and Systems*, 2:45–48, 1996.

[88] M. V. Wickerhauser. *Adapted Wavelet Analysis: From Theory to Software*. IEEE Press, 1994.

[89] Y.Ñ. P. Xiao Ping Zhang, M. Desai. Orthogonal complex filter banks ans wavelts: Some propierties and design. *IEEE Trans. on Signal Processing*, 47(4), April 1999.

[90] A. A. Xuedong Huang and H.-W. Hon. *Spoken Lenguage Processing: A Guide to Theory, Algorithm, and System Development*. 2001.

[91] S. S. e. a. Y. Chisaki, H.Ñakashima. A pitch detection method based on continuous wavelet transform for harmonic signal. *Acoust. Sci. and Tech.*, 24(1), 2003.

[92] A. M. y P. Stell. *Otorrinolaringología clínica*. Publicaciones Médicas, ESPAXS s.a., 1985.

[93] J. R. y S. Cercá. *Análisis de señales no estacionarias*. LLISO, 1996.

Apéndice A

Algoritmo de banco de filtros

En muchas aplicaciones, nunca se trata directamente con las funciones de escalamiento o las *wavelets*, en lugar de ello de trabaja con coeficientes que se convolucionan con las señales de entrada, esto coeficientes pueden ser tomados como filtros. El algoritmo de banco de filtros se basa en la teoría de multiresolución, y la consecución de dicho algoritmo es la continuación de la sección 3.2.2.

Si $\varphi(t)$ existe en V_0, también existe en V_1, el subespacio producto del *span* de $\psi(2t)$, entonces $\varphi(t)$ se puede expresar como una combinación lineal de versiones trasladadas de $\varphi(2t)$, de la forma

$$\sum_n h(n)\sqrt{2}\varphi(2t-n), \quad n \in \mathbb{Z} \tag{A.1}$$

donde los coeficientes $h(n)$ son una secuencia de números denominados coeficientes de la función de escalamiento o filtro de escalamiento, y el término $\sqrt{2}$ mantiene la norma de la función de escalamiento cuando la escala es de valor 2. En un espacio de multiresolución existe una base ortonormal para $\mathbf{L}^2(\mathbb{R})$

$$\psi_{j,n}(t) = 2^{-j/2}\psi(2^{-j}t-n), \quad j,n \in \mathbb{Z} \tag{A.2}$$

denominadas las *bases wavelet* tal que $\psi_{j,n}(t), n \in \mathbb{Z}$ es una base ortonormal para W_j. En efecto, la *wavelet* ψ puede ser construida en términos de la función de escalamiento φ por [18]

$$\psi(t) = \sqrt{2}\sum_{n\in\mathbb{Z}}(-1)^n h_{-n+1}\varphi(2t-n) \tag{A.3}$$

por lo tanto se tiene la siguiente representación *wavelet* discreta para $f \in \mathbf{L}^2(\mathbb{R})$:

$$f = \sum_{n\in\mathbb{Z}} c_{J,n}\varphi_{J,n} + \sum_{j\leq J}\sum_{n\in\mathbb{Z}} d_{j,n}\psi_{j,n} \tag{A.4}$$

donde los coeficientes de escalamiento son $c_{J,n} = \int_{-\infty}^{\infty} f(t)\varphi_{J,n}(t)dt$, y los coeficientes *wavelet* $d_{j,n} = \int_{-\infty}^{\infty} f(t)\psi_{j,n}(t); j \leq J$. Aquí $\sum_{n\in\mathbb{Z}} c_{J,n}\varphi_{J,n}$ es la aproximación de la función f a la resolución J. $\sum_{j\leq J}\sum_{n\in\mathbb{Z}} d_{j,n}\psi_{j,n}$, representa los detalles en términos de versiones Wavelet trasladas y dilatadas. Se puede mostrar que las siguientes relaciones se mantienen

$$c_{j,n} = \sum_{l\in\mathbb{Z}} h_{l-2n}c_{j-1,l}; \quad j,n \in \mathbb{Z} \tag{A.5}$$

$$d_{j,n} = \sum_{l\in\mathbb{Z}} g_{l-2n}c_{j-1,l}; \quad j,n \in \mathbb{Z} \tag{A.6}$$

$$c_{j-1,l} = \sum_{l\in\mathbb{Z}} h_{n-2l}c_{j,l} + \sum_{l\in\mathbb{Z}} g_{n-2l}d_{j,l}; \quad j,n \in \mathbb{Z} \tag{A.7}$$

donde $g_n = (-1)^n h_{-n+1}$. Se ve que los coeficientes de escalamiento y los coeficientes *Wavelet* pueden ser calculados recursivamente, ver fig. (A.1). h es el filtro pasa-bajas asociado a la función de escalamiento y g es el filtro pasa-altas asociado a la *Wavelet* madre. El algoritmo de la transformada *Wavelet* discreta puede ser usado para calcular los coeficientes [56].

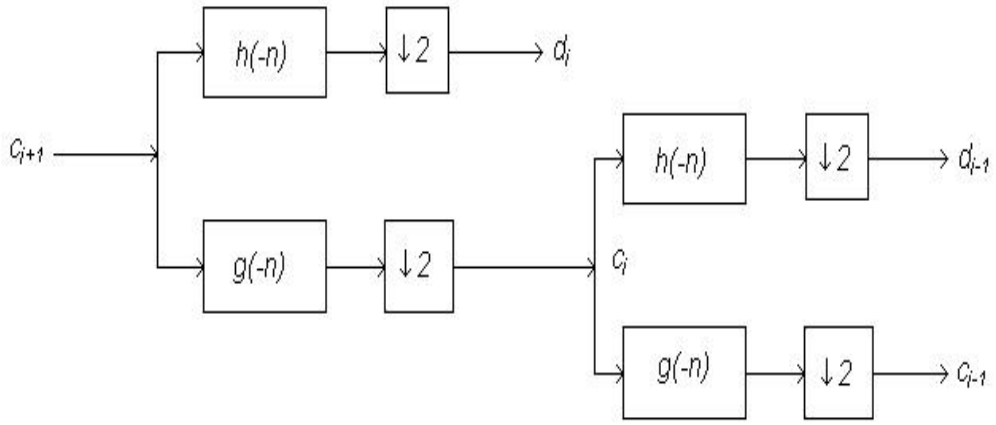

Figura A.1: Dos etapas del árbol de análisis de dos bandas

Apéndice B

Estimación en intervalos cortos de tiempo

B.1. Segmentación de los tipos de voz sonora/sorda usando childers

En primera instancia, es necesario enunciar el procedimiento para detección de voz/no voz, a partir del cual, se crea un factor discriminante respecto de los segmentos de señal a procesar (ya que se asume información irrelevante en intervalos no hablados [33]). Dicho procedimiento, utiliza como criterios de decisión [19]:

- El primer coeficiente de reflexión definido como:

$$R_1 = \frac{\sum_{n=1}^{N} s(n)s(n)}{\sum_{n=1}^{N-1} s(n)s(n+1)} \tag{B.1}$$

  Donde, N es número de muestras por ventana de análisis, al tiempo que $s(n)$ corresponde con la muestra de voz actual, y

- La señal residual, o aquella equivalente al error de predicción lineal (LPC - obtenida por el método de covarianza [86]), tomando como base, que la mayoría de características glotales se hallan contenidas en la misma [33].

Para los cuales se asume, un segmento hablado como aquel que posee un primer coeficiente de reflexión mayor a 0.2, en la misma medida que energía para error de predicción, superior a tres veces el valor de umbral (10^7) [19].

Por tanto se genera una secuencia binaria, en la cual, cada dígito se refiere a la decisión tomada sobre determinada ventana de señal, y donde además los patrones de tipo 101 y 010 (para 1 denotando segmentos de voz), se corrigen por sucesiones 111 y 000 respectivamente [19].

B.2. Estimación de la frecuencia fundamental usando Childers

En este punto, la señal residual obtenida con base en el proceso de predicción lineal, se integra para facilitar extracción de sus características glotales [34], a partir de un filtro que posee la forma:

$$H(z) = \frac{1}{1 - z^{-1}} \tag{B.2}$$

Posteriormente, este residuo integrado se utiliza como excitación para un filtro de paso bajo, con la siguiente relación de transferencia en dominio discreto [33]:

$$B(z) = B_1(z)B_2(z) = \left(\frac{1 - z^{-1}}{1 - 0{,}99z^{-1}}\right)\left(\frac{1 - z^{-1}}{(1 - 0{,}9z^{-1})\,(1 - 0{,}7z^{-1})}\right) \tag{B.3}$$

Donde, $B_1(z)$ pretende emular la inclinación espectral requerida para recuperar los pulsos glotales, al igual que $B_2(z)$ equivale a un filtro de fase cero, utilizado para reducir tanto ruidos de alta frecuencia, como derivas en baja escala [33].

Al procedimiento anterior, se le denomina de ”Filtrado inverso”, y genera una señal con alto grado de diferenciación entre pulsos glóticos.

Para la determinación de la frecuencia fundamental a partir del residuos del análisis LPC se usa el ceptrum [19]:

- Inicialmente, se aplica análisis Cepstral [85], sobre el segmento de señal en entrada.
- Posteriormente, se obtiene un pico de referencia, como el valor mayor para la secuencia obtenida en el tiempo.
- Luego, se calcula un pico de comparación, dentro del conjunto de eventos ubicados a la izquierda del valor derivado en el paso anterior.
- Finalmente, si el valor para dicho pico de comparación, es superior al 70 % del inicialmente calculado, se asignará como periodo fundamental. En caso contrario, se tomará como válida la decisión inicial.

Apéndice C

Wavelets spline

Introduciremos una familia de funciones, las *bases spline*, o *B-splines*, las cuales juegan un papel muy importante en variados campos de la ciencia. La funciones B-spline B_m, $m \geq 0$ son definidas recursivamente de la forma [15],

$$B_0(t)1_{[0,1]}(t) \tag{C.1}$$

y para, $m \geq 1$,

$$B_{m+1}(t) = (B_0 * B_m)(t) \tag{C.2}$$

Lo que equivale a decir que una función spline B de grado m puede ser calculada por la convolución consigo misma $m+1$ veces [15]. Donde su transformada de Fourier corresponde a [55],

$$\hat{B}(\omega) = \left(\frac{sin\frac{\omega}{2}}{\frac{\omega}{2}}\right)^{m+1} e^{-i\frac{\epsilon\omega}{2}} \tag{C.3}$$

Mediante la anterior generalización se obtiene una spline de orden arbitrario N diferenciable $N-1$ veces y con un soporte compacto de $N+1$. Dichas funciones poseen excelentes propiedades matemáticas, pero no son ortogonales respecto a la traslación. Si dicho sistema es ortogonalizado, el soporte compacto de vuelve infinito generando así el sistema *wavelet Battle-Lemarié* [17].

Para obtener la función de escalamiento ϕ del análisis de multiresolución $\{V_j\}_{j\in\mathbb{Z}}$ para el sistema Battle-Lemarié se aplica [55],

$$\hat{\varphi}(\omega) = \frac{e^{-i\frac{\epsilon\omega}{2}}}{\omega^{m+1}\sqrt{S_{2m+2}(\omega)}} \tag{C.4}$$

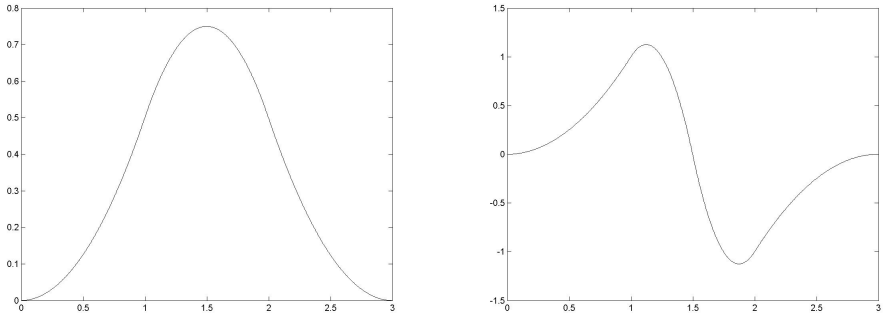

Figura C.1: Función de escalamiento y función *wavelet* (Análisis)

con

$$S_n(\omega) \sum_{k=-\infty}^{\infty} \frac{1}{(\omega + 2k\pi)^n} \tag{C.5}$$

El requerimiento de ortogonalidad de las expansiones *wavelet* respecto a las escalas y la traslación da como resultado ecuaciones de diseño complicadas y los filtros de síntesis y análisis podrían dar de fase no lineal, tal inconveniente se soluciona usando *wavelets* de base bi-ortogonal. Para sistemas ortogonales se cumplen las siguientes relaciones entre los filtros de análisis h, g y filtros de reconstrucción $\tilde{h}, \tilde{g}$: $\tilde{h}(n) = h(-n)$, $\tilde{g}(n) = g(-n)$. La familia de *wavelets* de base bi-ortogonal se obtienen al relajar la anterior condición permitiendo diferentes filtros para análisis y reconstrucción [17], dando como resultado la siguiente relación entre dichos filtros,

$$\tilde{g}(n) = (-1)^n h(1-n), \qquad \tilde{g}(n) = (-1)^n h(1-n) \tag{C.6}$$

En las figuras (C.1) se muestran las formas de la *wavelet* madre usadas para la descomposición (izquierda) y la reconstrucción (derecha). En (C.2) se aprecian las formas de onda para las funciones de escalamiento. Las anteriores formas de onda son usadas en la transformada *wavelet* bi-ortogonal, y corresponden a la *spline* cúbica.

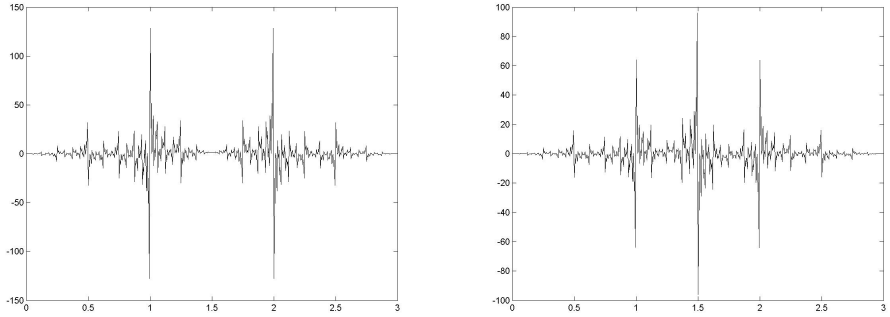

Figura C.2: Función de escalamiento y función *wavelet* (Reconstrucción)