UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA

FACULTAD DE INGENIERÍA

CODIFICACIÓN DE VOZ BASADA EN ANÁLISIS CEPSTRAL, ANÁLISIS WAVELET Y ESCALAS PERCEPTUALES DE FRECUENCIA

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN INGENIERÍA

MAESTRÍA EN INGENIERÍA ELÉCTRICA
PROCESAMIENTO DIGITAL DE SEÑALES

P R E S E N T A :

ING. FRANCISCO JAVIER AYALA SÁNCHEZ

T U T O R :

DR. JOSÉ ABEL HERRERA CAMACHO

JULIO DE 2011.

JURADO ASIGNADO:

Presidente: Dr. Escalante Ramírez Boris

Secretario: Dra. Medina Gómez Lucia

Vocal: Dr. Herrera Camacho José Abel

1er Suplente: Dr. Psenicka Bohumil

2do. Suplente: Dr. Orduña Bustamante Felipe

Facultad de Ingeniería, Ciudad Universitaria.

TUTOR DE TESIS:

Dr. José Abel Herrera Camacho

_____

FIRMA

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Los medios de comunicación por voz se han incrementado gracias a la evolución de las comunicaciones digitales. La compresión de señales de voz se ha extendido más allá de los sistemas de telefonía celular llegando a ser parte de la tecnología Voz Sobre IP (VoIP) utilizada en diversas aplicaciones de internet como las dedicadas a ofrecer el servicio de telefonía y las de mensajería instantánea.

Los codificadores de voz más utilizados son WMA voice (Windows Media Audio Voice) desarrollado por Microsoft; Speex, de desarrollo libre de patente y diseñado para VoIP principalmente y GSM 06.10 utilizado en telefonía celular. WMA y Speex compiten en el área de la compresión en VoIP junto con ACELP (Algebraic Code Excited Linear Prediction), el cual es un codificador que ha sido utilizado como parte de otros codificadores estandarizados.

Los codificadores mencionados se aplican en los sistemas de mensajería instantánea, en servicios de telefonía por internet basados en la tecnología VoIP, en consolas de videojuegos con interacción en línea entre los jugadores, en telefonía celular, entre otras. Las aplicaciones son diversas y los codificadores de voz de baja tasa de bits presentan una calidad regular en la voz, que aún no permite tener una experiencia satisfactoria en conversaciones a través de estos sistemas.

La técnica de procesamiento comúnmente utilizada en los codificadores antes mencionados es la Predicción Lineal. Toman alguna variante del sistema CELP como base de su funcionamiento. En la historia de la compresión de voz se han estudiado técnicas alternativas como el uso del cepstrum como base de funcionamiento.

En la década de los años 80 existía una gran cantidad de publicaciones sobre codificadores de voz basados tanto en el cepstrum y en predicción lineal (LP). Ambas técnicas competían fuertemente mas LP tomó ventaja dada su baja complejidad computacional pese al buen desempeño del análisis cepstral escalado en la frecuencia (warped cepstrum). Este último es costoso computacionalmente, situación poco propicia para la tecnología de la época.

La codificación basada en el cepstrum se dejó de tomar en cuenta y mientras mejoraban los codificadores LP, emergió una nueva técnica de procesamiento: la transformada wavelet discreta (DWT), que al ser combinada con algunas técnicas utilizadas en los sistemas LP para mejorar la calidad de la voz, y al ser utilizada en los modelos de codificación basados en cepstrum ofrece un nuevo modelo de codificación con buen desempeño y técnicas alternas a las tradicionales.

Para desarrollar el sistema de codificación de voz que se presenta en esta tesis, se retoma un codificador basado en mel-cepstrum y el filtro de síntesis MLSA. Se experimenta tanto con la escala mel y la escala Bark para adoptar la que ofrezca mejores resultados, en caso de haberlos. Se elimina el cálculo del periodo de altura tonal (pitch) y la generación de la señal de excitación mediante ruido blanco e impulsos periódicos. En su lugar, se utiliza el filtro MLSA inverso, el cual genera la señal de análisis. Ésta última se procesa con la DWT, para con ello extraer información necesaria para generar la señal se excitación.

El codificador de voz resultante es un sistema completo; es decir, contiene todas las etapas de un codificador estándar, incluyendo la cuantización de los parámetros y la asignación de bits. Este codificador, como se plantea más adelante, tiene buen desempeño a baja tasa de bits, comparado con el de los codificadores

mencionados antes. Cabe destacar que el diseño no es muy complejo en cuanto a etapas de procesamiento, contiene sólo los elementos fundamentales para ser un codificador, lo que le permite ser escalable. Lo anterior significa que se le pueden agregar etapas que mejoren su desempeño, sin incrementar la tasa de bits; se puede adaptar como codificador de aplicaciones muy diversas, ya sea, telefonía celular o VoIP.

La codificación o compresión de señales de voz es un procedimiento para representar una señal digitalizada con la menor cantidad posible de bits, asegurando una calidad razonable en la señal. El incremento en la demanda de medios de comunicación por voz ha hecho madurar la codificación de voz e incremetar el interés en su estudio y su estandarización. Actualmente constituye un campo de estudio muy importante en el procesamiento de señales.

La codificación de voz se lleva a cabo mediante un conjunto de operaciones o pasos que conforman un algoritmo. Un algoritmo es un conjunto de instrucciones bien definidas, ordenadas y finitas que permiten procesar un valor o un conjunto de valores de entrada para producir un valor o un conjunto de valores como salida.

1.1. Propiedades Deseables de un Codificador de Voz

El objetivo de la compresión de voz es minimizar la tasa de bits para una calidad perceptual particular o maximizar la calidad perceptual para una tasa de bits particular. Para seleccionar la tasa de bits se debe tomar en cuenta el costo de transmisión o de almacenamiento, el costo de compresión de la señal y los requerimientos de calidad de voz. En la mayoría de los codificadores hay diferencia entre la señal de salida y la original, debido a la pérdida de información redundante y de la precisión con la que se representa a los parámetros de la señal. Las propiedades deseables de un codificador de voz son [5]:

(a) Baja tasa de bits. Mientras menor sea la tasa de bits del flujo de datos, menor será el ancho de banda requerido para transmitir. Esta característica entra en conflicto con la calidad de la señal de salida.

(b) Alta calidad de la voz. La voz decodificada debe tener una calidad aceptable para la aplicación requerida. En la calidad de la voz se evalúa perceptualmente la inteligibilidad, la naturalidad y la facilidad para reconocer al hablante.

(c) Robustez en diferentes hablantes e idiomas. El sistema debe ser capaz de modelar de manera adecuada la voz de diferentes hablantes (hombres, mujeres, niños, etc.) y diferentes idiomas.

(d) Buen desempeño con señales diferentes a la voz. Aunque los codificadores de voz están diseñados para procesar, generalmente, sólo señales de voz, es deseable que puedan reconstruir otro tipo de señales, como tonos en el caso de telefonía, o música, aunque no lo hagan con la misma calidad que con la voz o por lo menos que no generen distorsiones molestas al procesar este tipo de señales.

(e) Baja complejidad computacional y uso de poca memoria. Con el propósito de que la implementación del codificador sea factible, los costos deben ser bajos; esto incluye la cantidad de memoria necesaria para su operación y la demanda computacional.

Los codificadores de voz se pueden clasificar por su tasa de bits [5]:

Tabla 1.1: *Clasificación de codificadores de voz de acuerdo a la tasa de bits*

Categoría	Rango de tasa de bits
Alta tasa de bits	$>$ 15 kbps
Mediana tasa de bits	5 a 15 kbps
Baja tasa de bits	2 a 5 kbps
Muy baja tasa de bits	$<$ 2 kbps

Además de la clasificación por tasa de bits, los codificadores pueden ser clasificados por las técnicas de codificación. Los codificadores de forma de onda intentan conservar la forma de onda de la señal original. La *relación señal ruido* (SNR, Signal-to-Noise Ratio) puede ser utilizada para medir la calidad de este tipo de codificadores. Los codificadores paramétricos generan un modelo de la señal original para extraer parámetros con los cuales se recontruye la señal decodificada. Este tipo de codificadores no intentan conservar la forma de onda original, por lo que la SNR no es útil para medir la calidad, la cual se mide con métodos de evaluación basados en la percepción. Los codificadores paramétricos trabajan con señales específicas, los modelos desarrollados están enfocados al procesamiento de señales de voz, lo que los hace poco eficientes con otras señales. Una combinación de las cualidades de un codificador de forma de onda con las de un codificador paramétrico forma un codificador híbrido.

Han existido diferentes codificadores de voz basados en análisis LPC (Linear Predictive Coding) desde el codificador LPC simple hasta el CELP (Code Excited Linear Prediction) utilizado en el estándar GSM [5]. LPC ha sido modificado para trabajar en escalas perceptuales y así disminuir la tasa de bits dada una calidad perceptual dando origen a WLPC (Warped LPC). Tales escalas perceptuales son la escala *mel* y la escala *Bark* de frecuencia.

Además de LPC, se ha utlizado el *análisis cepstral* en los codificadores de voz, el cual requiere una cantidad mayor de parámetros que un sistema basado en LPC. Sin embargo, se logra reducir dicha cantidad de parámetros al trabajar sobre una escala perceptual, mejorando la calidad de la voz.

1.2. Objetivos

Existe un codificador de voz basado en el cepstrum logarítmico en escala mel que utiliza el filtro MLSA para generar la señal sintética [16]. El objetivo es obtener una versión moderna del codificador basado en este principio, que tenga buen desempeño al recuperar señales de voz mezcladas con otro tipo de sonidos. Se introducirá el uso de señales residuales para generar la señal de excitación del filtro MLSA y no utilizar más la altura tonal (pitch) y las señales generadas con un tren de impulsos y ruido blanco.

Experimentar con el uso de la escala Bark, además de la escala mel, y evaluar las diferencias perceptuales entre las señales sintéticas generadas con ambas escalas y con ello concluir si alguna escala mejora la calidad de las señales.

Implementar un sistema de cancelación de ruido para mejorar el desempeño del codificador cuando en el entorno de uso exista ruido. Aún cuando se pretende que el codificador de voz sea capaz de sintetizar cualquier señal a parte de la voz, el uso de dicho sistema es opciona,l pero puede incorporarse para mejorar la experiencia de comunicación en ambientes ruidosos.

1.3. Definición del Problema

El sistema de codificación basado en el cepstrum en escala mel había sido planteado bajo un esquema básico en el que la señal de excitación del filtro de síntesis (MLSA) era generada con un tren de impulsos, ruido blanco y el valor de la altura tonal de acuerdo a la clasificación de tramas por su sonoridad. Esto da como resultado un codificador de muy baja tasa de bits dado que sólo utiliza un conjunto de parámetros cepstral para modelar la voz. Sin embargo la calidad de la señal sintética es baja, sin naturalidad, con dificultad para reconocer al hablante y presenta distorsiones molestas.

Se debe mejorar el sistema utilizando información residual tal como lo hacen los codificadores LPC, y para ello se plantea la obtención del filtro MLSA inverso con el cual se obtendrá información suficiente para generar la señal de excitación sin necesitar el valor de la altura tonal, un tren de impulsos ni ruido blanco.

El uso de la escala mel de frecuencia es muy usual en el análisis cepstral, por lo que se plantea el uso de la escala Bark como una posibilidad de mejoría de las señales sintéticas, lo cual deberá ser comprobado.

El sistema propuesto más adelante en esta tesis, busca tener la robustez de un sistema de codificación de telefonía, que al estar diseñado para codificar señales de voz principalmente, es capaz de reconstruir cualquier otro tipo de señal con una calidad aceptable, pero ante la incertidumbre de los posibles resultados, de la

aplicación que pudiera tener el sistema y el entorno de aplicación, se busca una alternativa que favorezca el desempeño del sistema propuesto la cual es contar con un sistema adaptativo de cancelación de ruido.

1.4. Alcance

En esta tesis, se muestra el desarrollo de un sistema codificador de voz completo que genera señales sintéticas de muy buena calidad, a tasas de bits bajas y medianas. Se alcanza una calidad similar a la que se obtiene en un codificador basado en LPC pero basando el modelo de análisis en el cepstrum de la señal en escala mel o Bark. Se realizaron pruebas MOS (Mean Opinion Score) para evaluar de manera subjetiva las señales resultantes del sistema codificador de voz con diferentes señales: voces masculinas, voces femeninas, voces con música, voces con ruido de la calle, etc.

El sistema de cancelación de ruido se implementa utilizando un filtro adaptativo y el algoritmo LMS. Al estar conformado de elementos básicos pero correctamente implementados, logra eliminar o atenuar toda información ajena a la voz. Cuando el ruido o la interferencia son de energía inferior a la de la voz, son eliminados completamente, si su energía es superior son atenuados de tal forma que sobresale la voz.

A continuación se presenta el resumen del contenido de la tesis.

Capítulo 1: Sirve como introducción, tiene como intención mostrar el alcance del proyecto y las implicaciones de la codificación de señales voz.

Capítulo 2: Explica el análisis cepstral y la metodología para extraer los coeficientes cepstral que posteriormente son utilizados como parámetros en el sistema de codificación.

Capítulo 3: Presenta una introducción a la transformada wavelet y un ejemplo de su aplicación en señales unidimensionales (señales de audio). La aplicación de ejemplo consiste en un sistema de compresión de señales de voz utilizando únicamente la transformada wavelet discreta.

Capítulo 4: Introduce el concepto del cambio de escala en el dominio de la frecuencia (frequency warping) al modo en el que se aplica en el sistema de compresión diseñado en el capítulo 5.

Capítulo 5: Presenta el diseño del decodificador propuesto en la tesis basado en el cepstrum, la transformada wavelet discreta y el filtro MLSA de síntesis.

Capítulo 6: Contiene un elemento adicional, útil para complementar al sistema del capítulo 5. Se trata de un sistema adaptativo de cancelación de ruido para la voz, que puede ser añadido al comienzo del codificador, o bien utilizado en alguna otra aplicación, como un sistema de reconocimiento de voz y así mejorar su desempeño, al cancelar la interferencia del ruido que se suma a la voz.

1.5. Aportaciones

El trabajo presente aporta lo siguiente:

Un sistema de codificación de voz de mediana tasa de bits basado en análisis cepstral y otros métodos no utilizados antes en este tipo de análisis, que fueron eficientes para resolver problemas específicos del sistema tales como el análisis wavelet y el uso de señales residuales.

Resultados basados en pruebas de percepción sobre el desempeño de las escalas Bark y mel en el sistema de codificación y una explicación completa sobre cómo hacer cambios de escala de frecuencia en señales y sistemas (frequency warping).

Un sistema de cancelación adaptativa de ruido del entorno en señales de voz que puede tener diversas aplicaciones no sólo en sistemas de codificación de voz, sino en sistemas de reconocimiento de voz cuyo desempeño decae con la presencia de ruido.

Capítulo 2

Análisis Cepstral

El análisis cepstral se aplica para extraer parámetros de la señal de voz original y transmitirlos como parte de la información de la señal comprimida. Los parámetros son obtenidos del cepstrum, el cual es el resultado de calcular la transformada inversa de Fourier IFT del logaritmo del espectro de la señal. El algoritmo para calcular el cepstrum es: señal $\rightarrow FT \rightarrow log(\mid FT \mid) \rightarrow IFT \rightarrow$ Cepstrum [27] [13].

La importancia del cepstrum es que permite separar las dos contribuciones del mecanismo de producción: la estructura fina del espectro, que caracteriza a la fuente de excitación, y la envolvente del espectro, que caracteriza al filtro asociado con el tracto vocal.

Si se denomina $x[n]$ a la señal de voz, la cual se deriva de la convolución entre la señal de excitación $g[n]$ y la respuesta al impulso del tracto vocal $h[n]$, y después se transforma al dominio de la frecuencia se tiene que:

$$X(\omega) = G(\omega) \cdot H(\omega) \tag{2.0.1}$$

después se obtiene el logaritmo del módulo de la expresión:

$$\log |X(\omega)| = \log |G(\omega) \cdot H(\omega)| \,, \tag{2.0.2}$$

se aplica la identidad logarítmica:

$$\log |X(\omega)| = \log |G(\omega)| + \log |H(\omega)| \tag{2.0.3}$$

y finalmente se calcula la IDFT, lo que resultará:

$$c\,(\tau) = IDFT\,(\log |X(\omega)|) = IDFT\,(\log |G(\omega)|) + IDFT\,(\log |H(\omega)|)\,. \tag{2.0.4}$$

$c[\tau]$ representa el cepstrum cuyo dominio es llamado quefrencia, y dado que es obtenido mediante la transformada inversa del dominio frecuencial, la quefrencia es una variable de dominio temporal. En la expresión del cepstrum se observa que aparecen como sumandos los componentes de los detalles y la envolvente espectral, por lo tanto se produce la deconvolución de los componentes de la señal de voz, bajo la suposición de que el filtro vocal (envolvente) corresponde principalmente con los elementos tempranos del cepstrum, y la señal de excitación con los elementos posteriores [4].

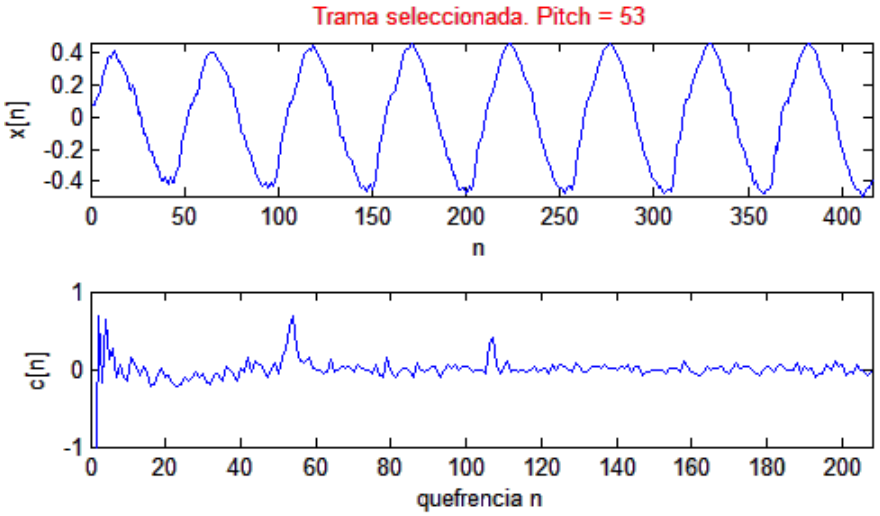

Figura 2.1: *a) Trama de una señal de voz y b) cepstrum de la trama.*

En la figura 2.1 se muestra el cepstrum de una trama de voz, donde los primeros coeficientes representan la respuesta al impulso del tracto vocal, y los picos equiespaciados y decrecientes que aparecen, corresponden a los componentes de la estructura armónica de la voz. La distancia entre dichos picos corresponde a la frecuencia fundamental de la señal (pitch).

Para extraer la envolvente del espectro, es necesario retener los coeficientes del cepstrum cercanos al origen, lo cual se logra haciendo un filtrado en el dominio cepstral, lo que equivale a multiplicar el cepstrum por una ventana que retenga sólo los coeficientes de bajas quefrencias [4]. Una vez extraidos los coeficientes, es posible obtener una buena aproximación a la envolvente espectral al calcular la transformada de Fourier sobre éstos [5].

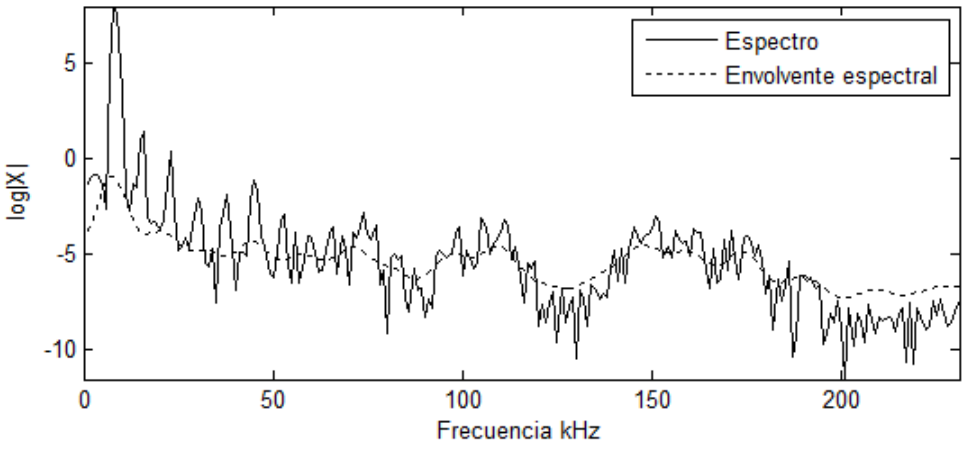

Figura 2.2: *Espectro de la trama de voz y su envolvente espectral.*

En la figura 2.2 se muestra el espectro de la señal de la figura anterior y la envolvente del espectro obtenida mediante la DFT de los primeros treinta coeficientes del cepstrum.

Capítulo 3

Compresión de Señales de Voz Utilizando la Transformada Wavelet Discreta

La transformada wavelet discreta (DWT) es una herramienta matemática que permite el análisis de señales de una manera alternativa a la transformda de Fourier (TF). La DWT puede entregar información temporal y frecuencial simultáneamente, mientras que la TF da una sola representación frecuencial. La importancia de la DWT en este trabajo reside en una de sus aplicaciones: la codificación de señales.

3.1. Desventaja del Análisis de Fourier ante las Funciones Wavelet

La transformada de Fourier (TF) expresa una señal en términos de combinaciones lineales de exponenciales complejas. Para una señal real se puede mostrar que la TF expresa la señal en términos de una combinación lineal de senos y cosenos [8].

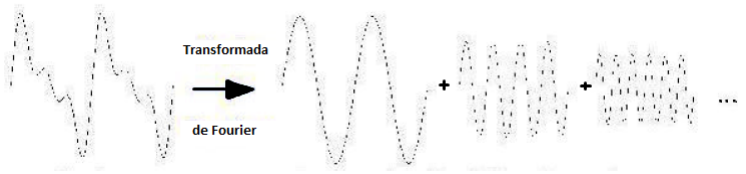

Figura 3.1: *Método de la trasformada de Fourier.*

En la figura 3.1 se observa que la salida de la TF esta constituida por una combinación de sinusoidales de diferentes frecuencias, tales frecuencias son las que conforman a la señal original. Además existe un cambio de dominio, la señal original está expresada en función del tiempo mientras que la señal transformada está expresada en función de la frecuencia.

La TF es una herramienta muy útil para conocer el contenido frecuencial de una señal, sin embargo presenta el inconveniente de no indicar la posición en el tiempo de cada frecuencia. Por ejemplo, si generamos una señal $x_1(n)$ senoidal de 1000 puntos de longitud donde la frecuencia de la señal $x_1(1:500)$ es $f_1 = 20Hz$ y la frecuencia de $x_1(501:1000)$ es $f_2 = 100Hz$ y comparamos su espectro con el de otra señal $x_2(n)$ cuya frecuencia para $x_2(1:500)$ es f_2 y para $x_2(501:1000)$ es f_1, observaremos que es igual en magnitud, pero con diferencias en la fase de los espectros.

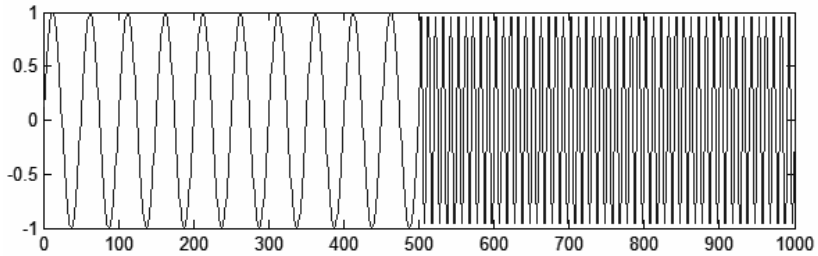

Figura 3.2: *Señal x_1*

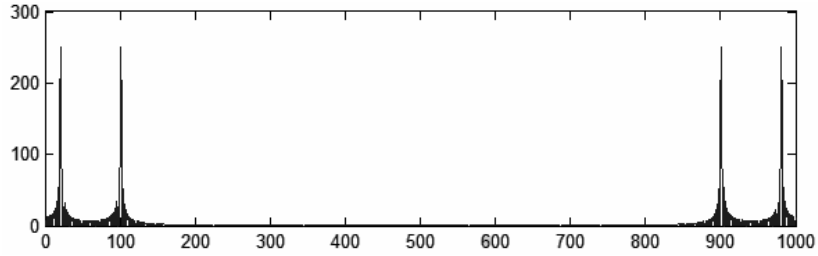

Figura 3.3: *FFT de x_1*

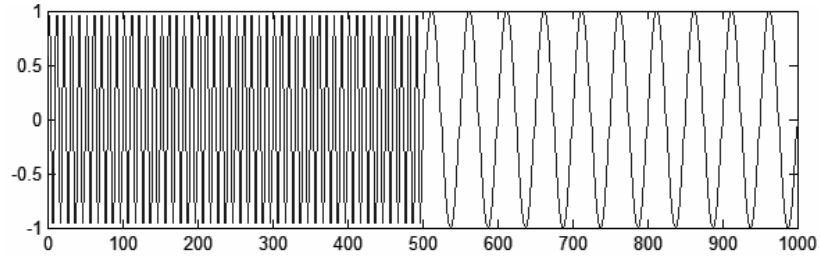

Figura 3.4: *Señal x_2*

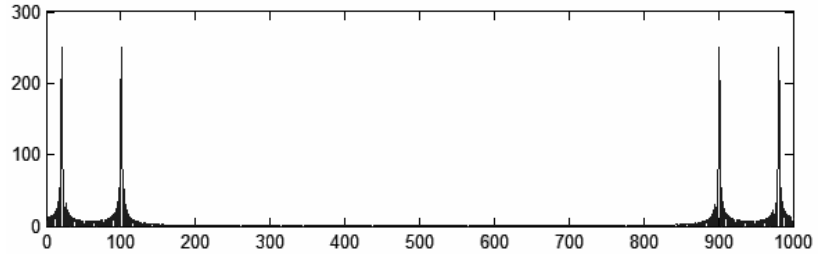

Figura 3.5: *FFT de x_2*

El ejemplo anterior muestra la desventaja del análisis de Fourier al no poder distinguir diferencias de magnitud espectral entre dos señales diferentes, con la misma información de frecuencias, en diferentes

ubicaciones en el tiempo. Por lo tanto la FT no es ideal para al análisis de señales no estacionarias en determinadas aplicaciones.

3.2. La Transformada Wavelet

Sea $\psi(t)$ una función real o compleja continua con las siguientes propiedades [19]:

(a) La integral de la función es cero. $\int_{\infty}^{\infty} \psi(t).d(t) = 0$

(b) Su cuadrado es integrable o, es de energía finita. $\int_{\infty}^{\infty} |\psi(t)|^2 < \infty$

Se llama mother wavelet a aquella función que satisface las dos propiedades. Existe una gran cantidad de funciones que satisfacen las dos propiedades, la más simple de ellas es la wavelet "Haar". Una familia de wavelets útiles para este análisis es la familia "Daubechies"la cual tiene wavelets db1 a db10 [8].

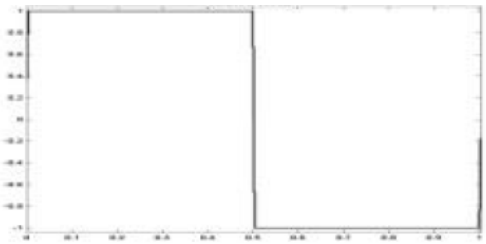

Figura 3.6: *Wavelet Haar.*

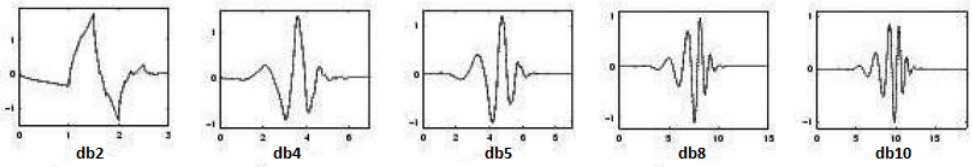

Figura 3.7: *Algunas wavelets Daubechies.*

3.2.1. Comparación de una Señal Senoidal con una Wavelet

La función wavelet es una onda de duración limitada y su promedio es cero. Las ondas senoidales no son de duración limitada, se extienden desde menos infinito a mas infinito. Las ondas senoidales son suaves y predecibles mientras que las wavelets tienden a ser irregulares y asimétricas.

El análisis de Fourier consiste en descomponer una señal en ondas senoidales de diferentes frecuencias. El análisis wavelet es la descomposición de una señal en versiones escaladas y desplazadas de la función wavelet [21].

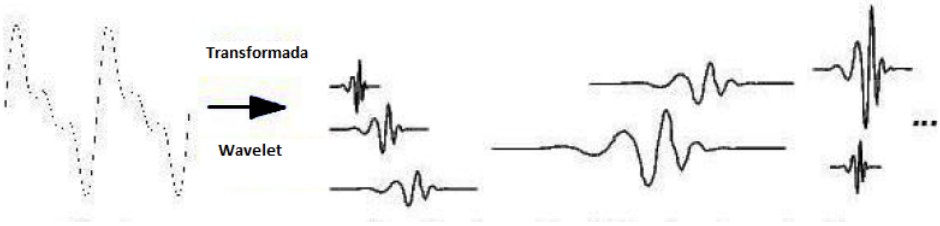

Figura 3.8: *Transformada wavelet, constituida por diferentes escalas y posiciones de la función wavelet.*

La posición de la wavelet nos proporciona la ubicación en el tiempo de información frecuencial de la señal. La escala de la función wavelet está relacionada con la frecuencia de la señal [19].

Baja escala significa una wavelet comprimida → cambios rápidos en la señal, detalles → altas frecuencias.

Alta escala significa una wavelet alargada → cambios lentos en la señal → bajas frecuencias [8].

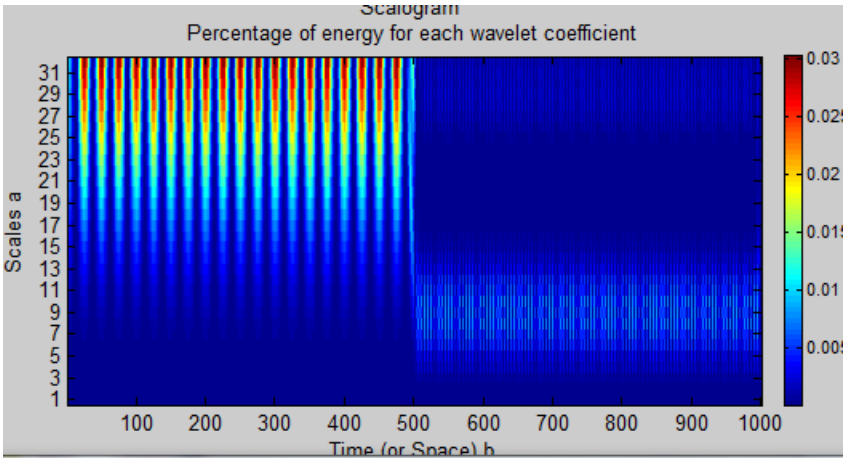

Figura 3.9: *Escalograma de la señal x_1 utilizando la wavelet Haar.*

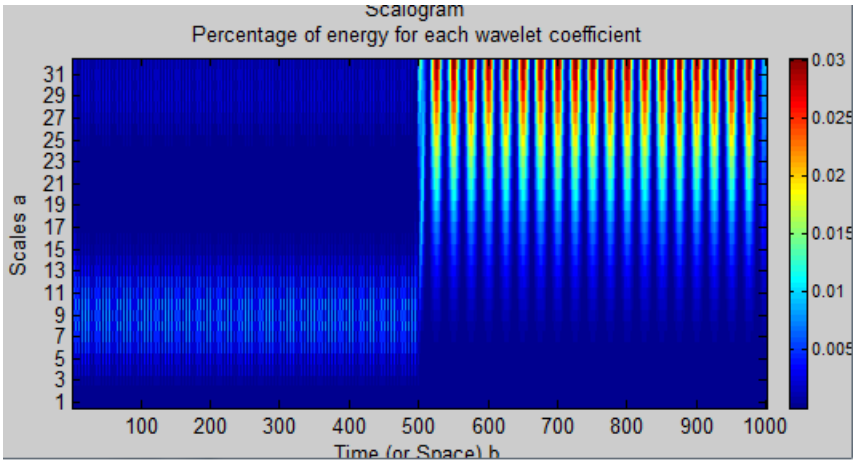

Figura 3.10: *Escalograma de la señal x_2 utilizando la wavelet Haar.*

En las gráficas 3.9 y 3.10 podemos observar que los escalogramas de las señales x_1 y x_2 son diferentes. Mediante escalas y desplazamientos se muestra información que corresponde a detalles y aproximaciones.

3.3. Compresión

Para el proceso de compresión, primero se aplica la transformada wavelet discreta en la cual, para evitar el crecimiento de la información, se aplica un submuestreo en la señal antes de aplicarle el filtro generado con los coeficientes de la mother wavelet seleccionada. Dicho submuestreo genera el efecto de aliasing, por lo que se aplica el concepto de filtros espejo en cuadratura. En cada nivel de descomposición se aplica un submuestreo y se pasa la señal por un filtro paso bajas, generado con los coeficientes de la función de escala $\phi(n)$ y por un filtro paso altas generado con los coeficientes de la función wavelet $\psi(n)$, con lo que se obtienen dos salidas, la primera llamada coeficientes de aproximación y la segunda coeficientes de detalles [19].

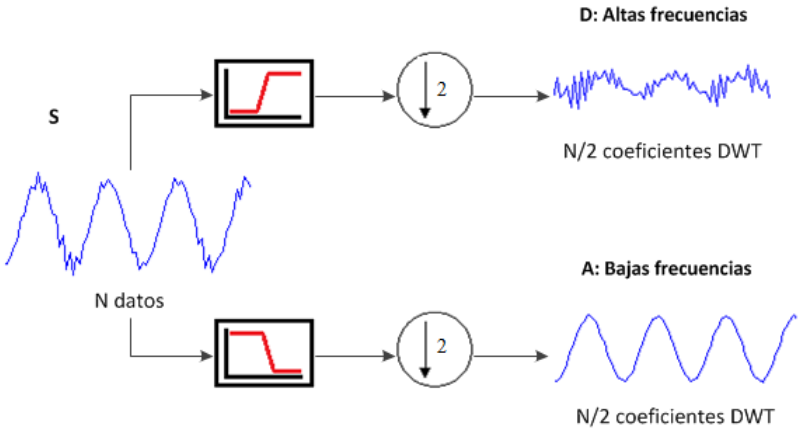

Figura 3.11: *Ejemplo de la obtención de los coeficientes de detalle y de aproximación del primer nivel de filtrado.*

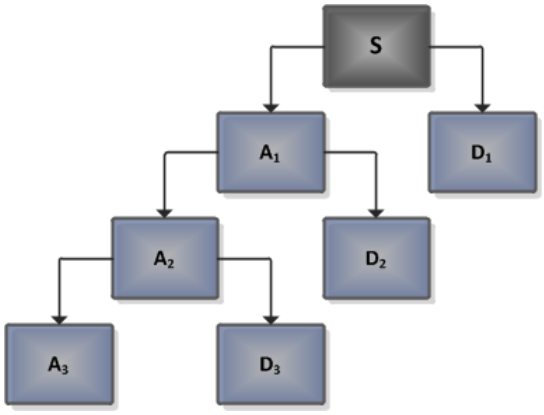

Figura 3.12: *Árbol de descomposición en diferentes niveles y la relación entre la señal y sus componentes* $S = A_3 + D_1 + D_2 + D_3$.

En el proceso de análisis wavelet la señal es descompuesta en señal de aproximación y señal de detalles. La señal de aproximación la conforman los componentes de baja frecuencia de la señal y la señal de detalles

la conforman los componentes de alta frecuencia. El proceso de descomposición se logra mediante iteraciones; una señal es dividida en varias señales de baja resolución (decimación diádica) en cada iteración [8].

Dada una señal de longitud N, se le aplica la DWT de L niveles de descomposición. El primer paso produce dos conjuntos de coeficientes: coeficientes de aproximación A1 y coeficientes de detalles D1. Ambos vectores son obtenidos al realizar la convolución de la señal con los coeficientes del filtro paso bajas para la aproximación, y con los coeficientes del filtro paso altas para los detalles. Este proceso va seguido de la decimación de los vectores.

En el segundo paso, los coeficientes de aproximación A1 son divididos en dos partes para repetir el procedimiento anterior, así se reemplaza la señal de entrada por A1 y se produce A2 y D2 y así sucesivamente. Después de estos pasos se obtiene un árbol de descomposición de L niveles. Para realizar el proceso de reconstrucción, se aplica la transformada inversa wavelet IDWT a los coeficientes de aproximación y detalles. Los vectores resultantes son posteriormente sobremuestreados y filtrados. El sobremuestreo se logra aplicando el *rellenado con ceros* para recuperar la longitud original de la señal al final de las iteraciones [19]. Para el diseño de los filtros se toma los coeficientes de la función wavelet seleccionada, que en el procesamiento de señales de voz, la wavelet db2 es utilizada.

Los coeficientes del filtro son [25]:

(a) Filtro de descomposición paso altas: $h_0(n)$={-0.4830,0.8365,-0.2241,-0.1294}

(b) Filtro de descomposición paso bajas: $h_1(n)$={-0.1294,0.2241,0.8365,0.4830}

(c) Filtro de reconstrucción paso altas: $h_2(n)$={-0.1294,-0.2241,0.8365, -0.4830}

(d) Filtro de reconstrucción paso bajas: $h_3(n)$={0.4830,0.8365,0.2241,-0.1294}

Donde la relación entre los filtros de descomposición y reconstrucción es $H_2(z) = H_1(-z)$ y $H_3(z) = -H_0(-z)$ dada la reconstrucción perfecta y la cancelación del efecto *aliasing*.

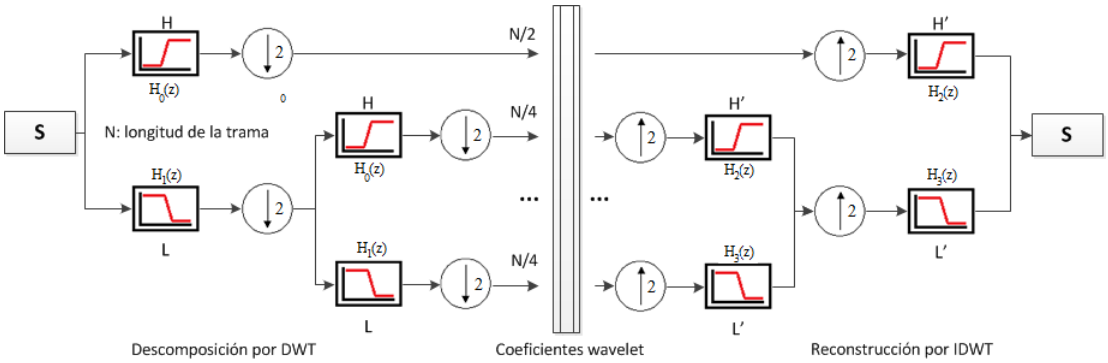

Figura 3.13: *Filtros espejo en cuadratura. Proceso de análisis y síntesis multinivel.*

La elección de la función wavelet ideal para la compresión requiere tomar en cuenta varios criterios, tales como minimizar el error de la varianza de la reconstrucción, maximizar la relación señal a ruido (SNR), que en el primer nivel de descomposición la energía de los coeficientes de aproximación sea la mayor, la correlación entre la wavelet y la señal sea mayor, etc.

3.3.1. Proceso de Compresión

El proceso de compresión consiste en truncar a cero los coeficientes de la DWT que estén por debajo de un umbral seleccionado (dicho umbral se selecciona en base a experimentos). Experimentos realizados muestran que el 90 % de los coeficientes tienen valor casi cero por lo que se consideran insignificantes y pueden ser descartados [8].

3.3.2. Codificación de los Coeficientes

Una vez que han sido eliminados los coeficientes insignificantes, se crean dos vectores, uno con los coeficientes que quedaron (sin los ceros) y otro vector con las posiciones y números de ceros para posteriormente, poder reconstruir los coeficientes con la posición original. Estos dos vectores son los que se transmiten, para que con ellos de realice la reconstrucción de la señal. La diferencia de estos coeficientes con los originales es que los que tenían valores muy pequeños, casi cero, ahora son cero, lo que generará un error.

En la figura 3.14 se muestra el diagrama correspondiente al proceso de compresión utilizando la transformada wavelet discreta.

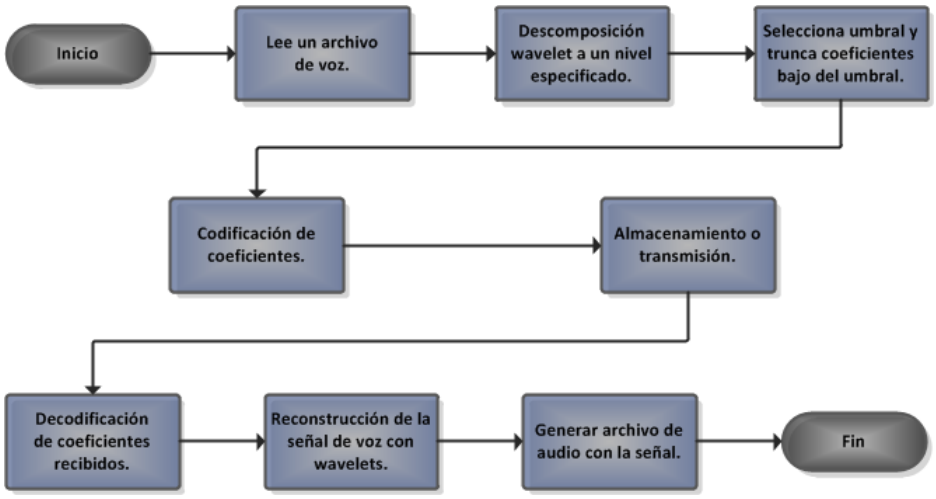

Figura 3.14: *Diagrama de compresión y reconstrucción de voz.*

3.4. Ejemplos

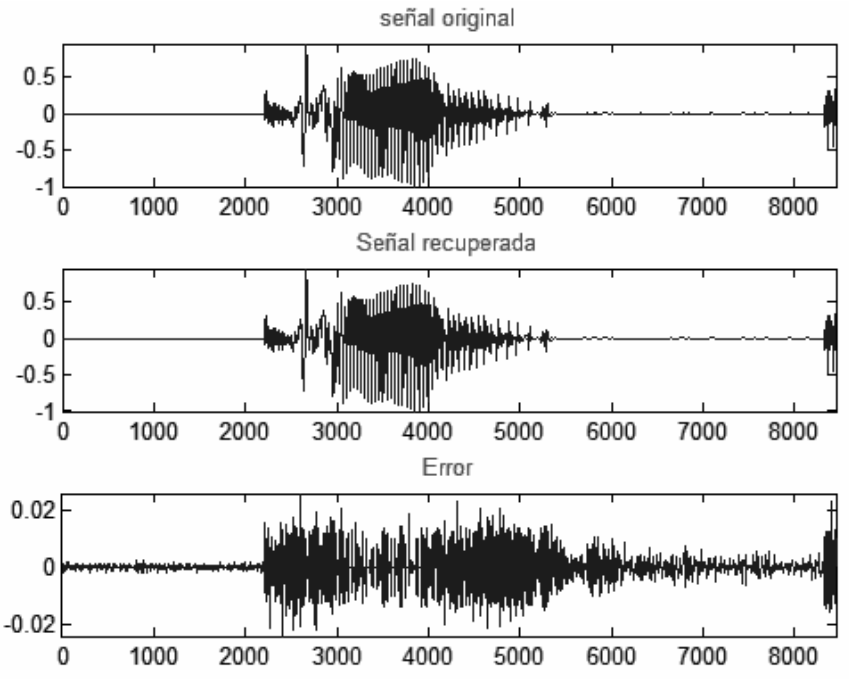

Figura 3.15: *Señal original y señal recuperada aplicando la wavelet Haar. Umbral=0.02*

En la figura 3.15 se utilizaron 5 niveles de descomposición, la wavelet Haar y un umbral igual a 0.02.
Tasa de compresión: 0.36. Porcentaje de ceros: 70.8 %.
Si aumentamos el umbral podemos comprimir más la señal sacrificando calidad en la recuperación.

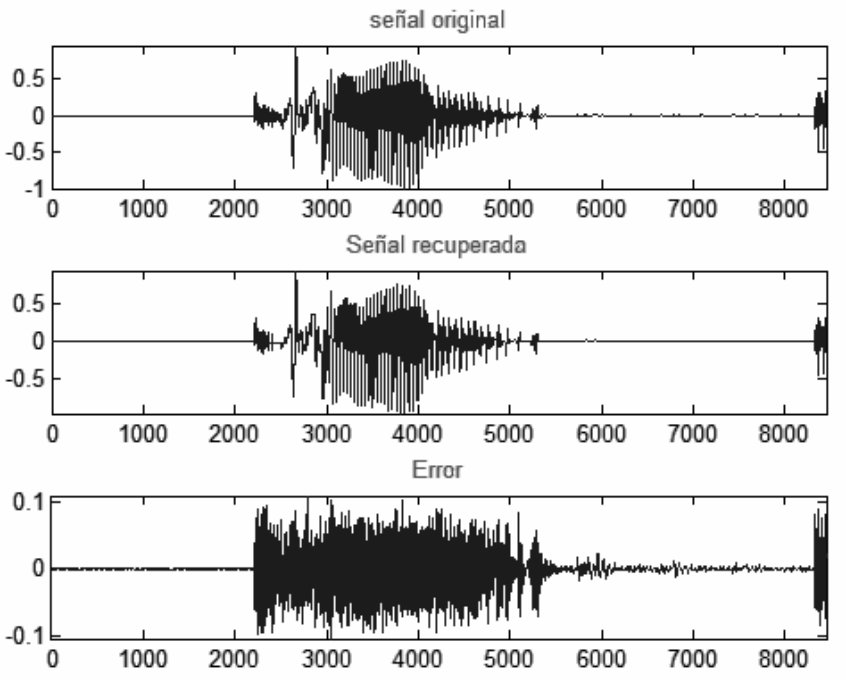

Figura 3.16: *Señal original y señal recuperada aplicando la wavelet Haar. Umbral=0.1*

La señal resultante, en la figura 3.16, es clara al ser escuchada aunque con un poco de ruido. Se observa la disminución en nitidez pero es aceptable. El umbral es igual a 0.1. Tasa de compresión: 0.18. Porcentaje de ceros: 86.6 %.

La gráfica de la figura 3.17 muestra la compresión de la señal utilizando la wavelet db2, con un umbral igual a 0.02.

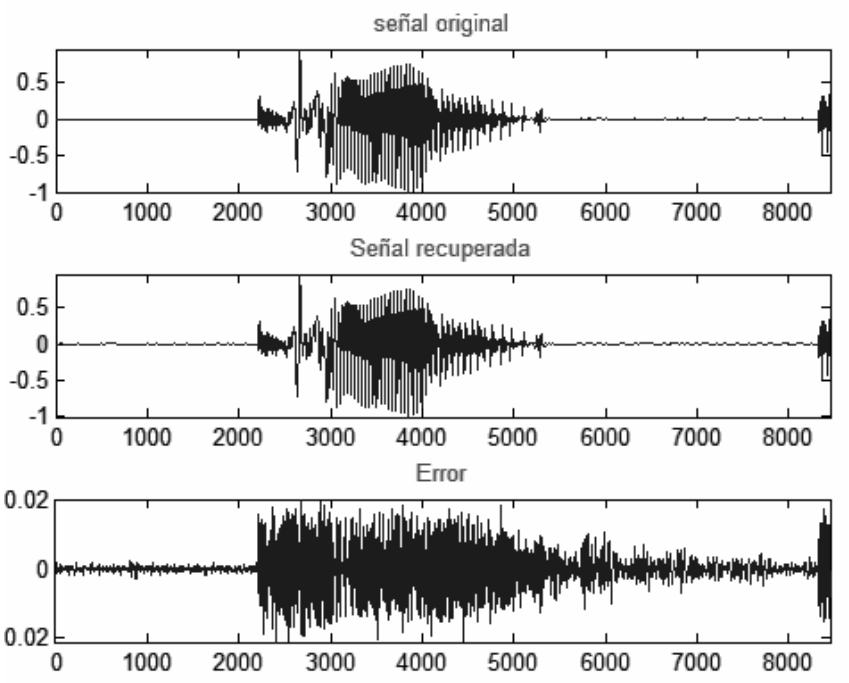

Figura 3.17: *Señal original y señal recuperada aplicando la wavelet db2. Umbral=0.02*

Tasa de compresión: 0.30. Porcentaje de ceros: 76.8 %.

Otro experimento con db10:

Tasa de compresión: 0.27. Porcentaje de ceros: 78.3 %.

En el sistema desarrollado en esta tesis se aplica la compresión con wavelets a la señal de salida del filtro MLSA inverso.

Capítulo 4

Cambio de Escala en el Dominio de la Frecuencia (Frequency Warping)

El sistema auditivo humano es un analizador de alta complejidad, el cual es un sistema no lineal, variante en el tiempo y adaptativo de diferentes maneras [6]. Por lo tanto, los modelos de percepción auditiva son necesariamente complejos y las técnicas requeridas para utilizar estos principios son complicadas. La técnica más común para representar el comportamiento del sistema auditivo humano es el uso de escalas de frecuencia que son no lineales y no uniformes, en relación con la escala en hertz (Hz). Ejemplos de dichas escalas son la escala mel, la escala Bark y la escala *ERB* (Ancho de banda rectangular equivalente) [6] [1]. El uso de estas escalas es deseable en sistemas de procesamiento de señales de audio, cuyos resultados se analizan con criterios de percepción. La escala *mel* y la escala *Bark* mejoran el desempeño del sistema de codificación de señales de voz [16] [1] [17] [9].

La escala de frecuencia convencional en el procesamiento de señales es lineal en relación con la escala en hertz, su resolución es uniforme desde la componente DC (corriente directa) hasta la frecuencia máxima delimitada por la frecuencia de Nyquist, es decir, la mitad de la frecuencia de muestreo fs. Este hecho se origina con la propiedad del *retardo unitario* $D(z) = z^{-1}$, que significa retrasar todas las muestras de la señal el mismo intervalo de tiempo discreto. Para lograr una escala no lineal, se recurre a una transformación bilineal, la cual se deriva de un mapeo conforme, dicho mapeo se obtiene a través de un filtro pasa todas (AP) de primer orden [1].

4.1. Mapeo Conforme

Un mapeo conforme es una función que conserva los ángulos. También es llamada transformación de conservación de ángulo. Es una transformación conforme $w = f(z)$. Los mapeos más comunes se realizan entre dominios pertenecientes al plano complejo. Una función analítica es conforme en cualquier punto que tenga derivada diferente a cero [14].

Sea f una función analítica en el dominio D y sea z_0 un punto en D. Si $f^{'}(z_0) \neq 0$, entonces f puede ser expresada en la forma

$$f(z) = f(z_0) + f^{'}(z_0)(z - z_0) + \eta(z)(z - z_0) \tag{4.1.1}$$

donde $\eta(z) \to 0$ mientras que $z \to z_0$. Si z tiene un valor cercano a z_0, entonces la transformación $w = f(z)$ tiene la aproximación lineal

$$S(z) = A + B(z - z_0) = Bz + A - Bz_0 \tag{4.1.2}$$

donde $A = f(z_0)$ y $B = f^{'}(z_0)$. Dado que $\eta(z) \to 0$ cuando $z \to z_0$, para puntos cercanos a z_0 la transformación $w = f(z)$ tiene un efecto muy similar al mapeo lineal $w = S(z)$. El mapeo lineal S provoca

una rotación del plano un ángulo $\alpha = \arg f'(z_0)$, que se refiere al argumento del valor complejo de z_0, seguida de un escalamiento con el factor $\left|f'(z_0)\right|$ y finalizando con la traslación por el vector $A - Bz_0$. Por consiguiente, el mapeo $w = S(z)$ conserva los ángulos en el punto z_0. El mapeo $w = f(z)$ también conserva los ángulos en z_0.

Sea $C : z(t) + iy(t)$, $-1 \leq t \leq 1$ una curva suave que pasa por el punto $z(0) = z_0$. Un vector T tangente a la curva C en el punto z_0 esta dado por

$$T = z'(0), \tag{4.1.3}$$

donde el número $z'(0)$ es expresado como vector.

El ángulo de inclinación de T con respecto al eje x es

$$\beta = \arg z'(0) \tag{4.1.4}$$

La imagen de C bajo el mapeo $w = f(z)$ es la curva K dada por la fórmula $K : w(t) = u(x(t), y(t)) + iv(x(t), y(t))$. El vector T^* tangente a K en el punto $w_0 = f(z_0)$ está dado por

$$T^* = w'(0) = f'(0)z'(0). \tag{4.1.5}$$

El ángulo de inclinación de T^*con respecto al eje positivo u es

$$\gamma = \arg f'(z_0) + \arg z'(0) = \alpha + \beta, \tag{4.1.6}$$

donde $\alpha = \arg f'(z_0)$. Por consiguiente, el efecto de la transformación $w = f(z)$ es el de girar el ángulo de inclinación del vector tangente T en z_0 la magnitud del ángulo $\alpha = \arg f'(z_0)$ para conocer el ángulo de inclinación del vector tangente T^* en w_0 [14].

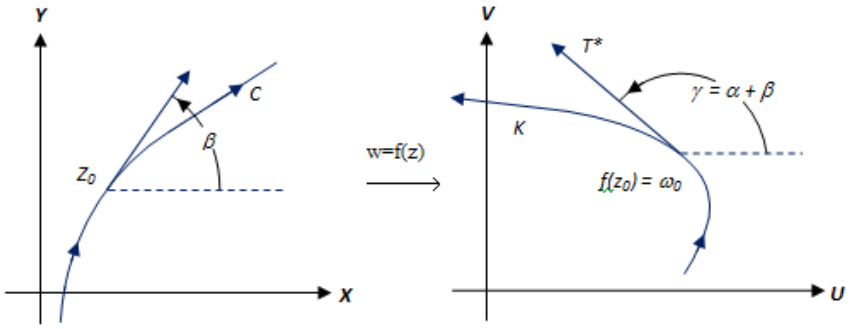

Figura 4.1: *Curvas suaves y sus tangentes en los puntos z_0 y w_0. $f'(z_0) \neq 0$.*

Se dice que el mapeo $w = f(z)$ es conservador de ángulo o conforme en z_0, si conserva los ángulos entre curvas tanto en magnitud como en orientación.

Sean C_1 y C_2 dos curvas suaves que pasan por z_0 con tagentes T_1 y T_2 respectivamente. Las curvas imágenes K_1 y K_2 que pasan por el punto $w_0 = f(z_0)$ tienen tangentes T_1^* y T_2^* respectivamente. De la ecuación (4.1.6) los ángulos de inclinación γ_1 y γ_2 de T_1^* y T_2^* están relacionados con β_1 y β_2 mediante las ecuaciones

$$\gamma_1 = \alpha + \beta_1 \tag{4.1.7}$$

y

$$\gamma_2 = \alpha + \beta_2 \tag{4.1.8}$$

donde $\alpha = \arg f'(z_0)$. Por lo tanto de las ecuaciones (4.1.7) y (4.1.8) se concluye que

$$\gamma_2 - \gamma_1 = \beta_2 - \beta_1 \tag{4.1.9}$$

El ángulo $\gamma_2 - \gamma_1$ de K_1 a K_2 es el mismo en magnitud y en orientación como el ángulo $\beta_2 - \beta_1$ de C_1 a C_2. En conclusión, el mapeo $w = f(z)$ es conforme en z_0 [14].

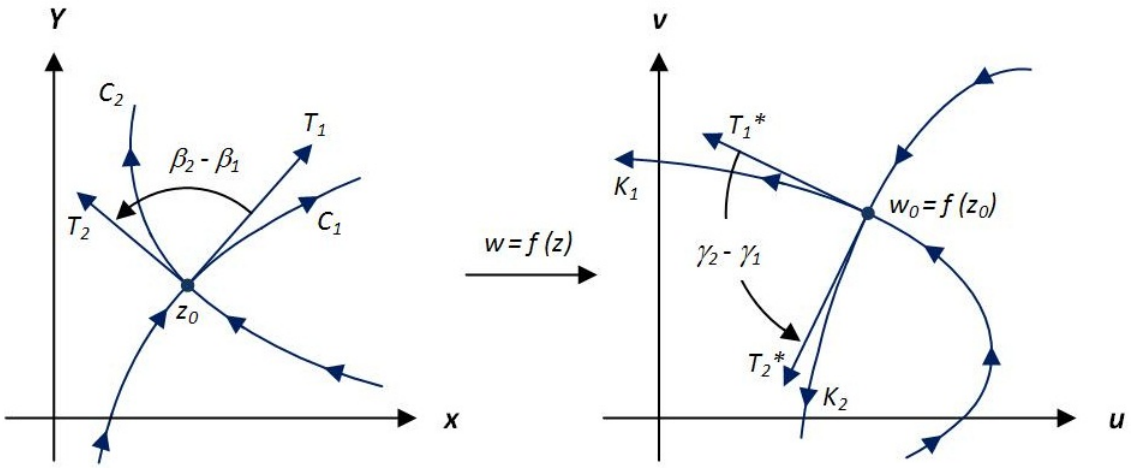

Figura 4.2: *Mapeo conforme. Conservación de ángulos entre ambas tangentes.*

4.2. Transformaciones Bilineales

Este tipo de mapeo es expresado convenientemente como el cociente de dos expresiones lineales. Existe una transformación bilineal única que mapea tres puntos distintos z_1, z_2 y z_3 a tres diferentes puntos w_1, w_2, y w_3 respectivamente. Entonces, con sólo tres puntos específicos y sus imágenes es posible determinar el mapeo para todo z y w. La fórmula implícita del mapeo bilineal es [1]

$$\frac{z - z_1}{z - z_3}\frac{z_2 - z_3}{z_2 - z_1} = \frac{w - w_1}{w - w_3}\frac{w_2 - w_3}{w_2 - w_1}. \tag{4.2.1}$$

Las transformaciones bilineales mapean círculos y líneas en otros círculos y líneas. En audio digital, donde ambos dominios son el plano z normalmente se requiere mapear el círculo unitario en sí mismo, con el mapeo de dc a dc ($z_1 = w_1 = 1$) y la frecuencia máxima a la frecuencia máxima ($z_2 = w_2 = -1$) [1]. Al hacer estas sustituciones en (4.2.1) la transformación queda con la forma

$$z = \frac{w + \alpha}{1 + \alpha w}, \tag{4.2.2}$$

donde

$$\alpha = \frac{w_3 - z_3}{1 - z_3 w_3} \tag{4.2.3}$$

ver demostración en Apéndice A.

La constante α provee un grado de libertad suficiente para poder mapear cualquier frecuencia ω, correspondiente al punto $e^{j\omega}$ del círculo unitario, a una nueva ubicación $a(\omega)$. El resto de las frecuencias serán reubicadas por consiguiente. El coeficiente pasa todas α determina la escala a la que se realizará el mapeo de la frecuencia lineal y su valor es de tipo real. La función inversa de $z(\alpha)$ se obtiene haciendo $z(-\alpha)$ intercambiando las variables z y w:

$$z = \frac{w + \alpha}{1 + \alpha w} \tag{4.2.4}$$

$$z + z\alpha w = w + \alpha \tag{4.2.5}$$

$$w\left(z\alpha - 1\right) = \alpha - z \tag{4.2.6}$$

$$w = \frac{z - \alpha}{1 - \alpha z} \tag{4.2.7}$$

El diagrama de polos y ceros de la ecuación (4.2.2) tomando en cuenta que α es un valor real es mostrado en la figura 4.3.

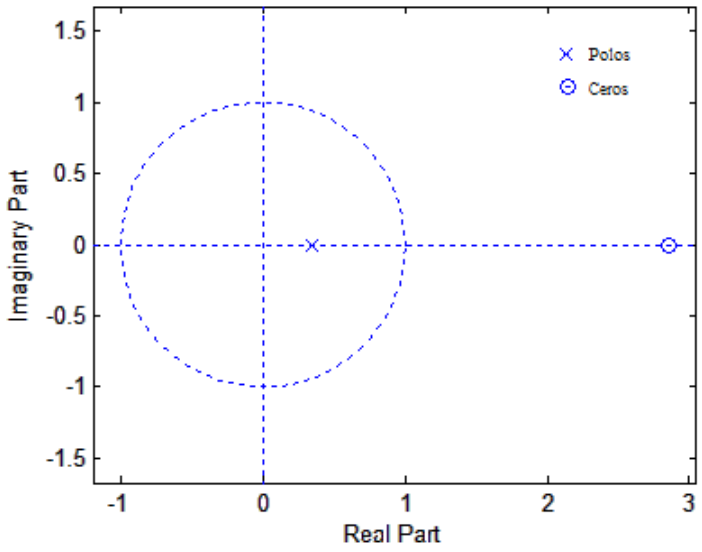

Figura 4.3: *Círculo unitario en el plano z del filtro pasa todas.*

Tomando en cuenta que al aplicar el cambio de escala de frecuencia a un sistema se sustituye cada retardo unitario z^{-1} por el filtro pasa todas, la forma en la que queda la ecuación (4.2.2) es

$$\tilde{z} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, |z| < 1. \tag{4.2.8}$$

4.3. Filtro Pasa Todas

La función de transferencia del filtro pasa todas está dada por la ecuación (4.2.2) [6]. Por definición, la magnitud de la respuesta en frecuencia es constante. La fase es variante de acuerdo al valor de α y se observa en la figura 4.4. Si $\alpha = 0$ la función de transferencia se reduce a un sólo retardo unitario con fase lineal y retardo de grupo constante.

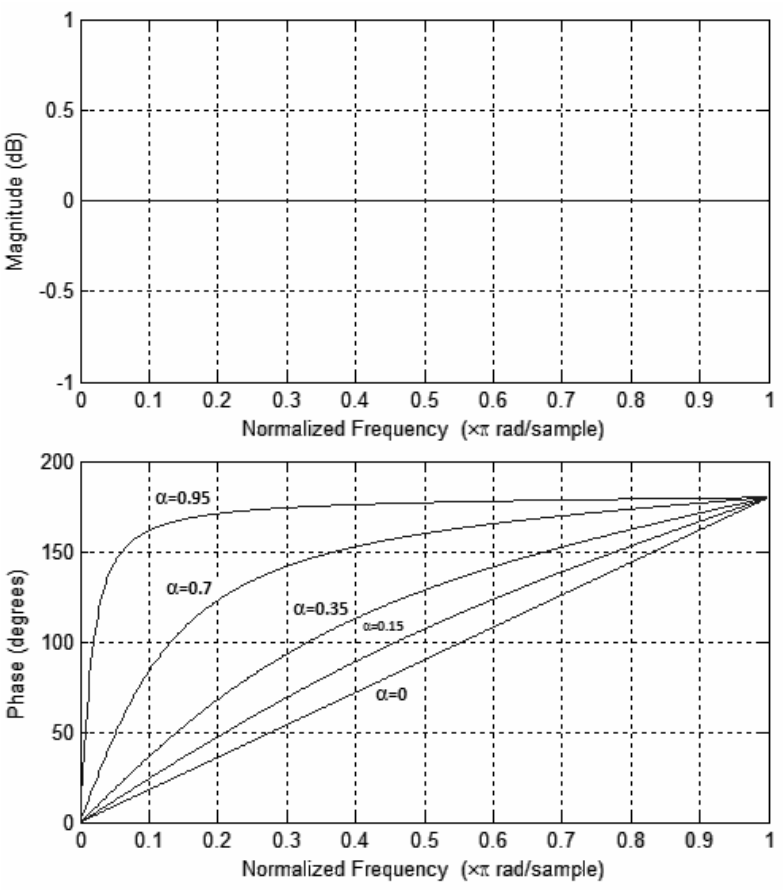

Figura 4.4: *Respuesta en frecuencia y en fase del filtro pasa todas con diferentes valores de* α.

El retardo de grupo (group delay) del filtro se muestra en figura 4.5. Por ejemplo, según la gráfica el retardo de grupo del filtro con $\alpha = 0.7$ es de casi 6 muestras en bajas frecuencias pero menos de 1 muestra a altas frecuencias.

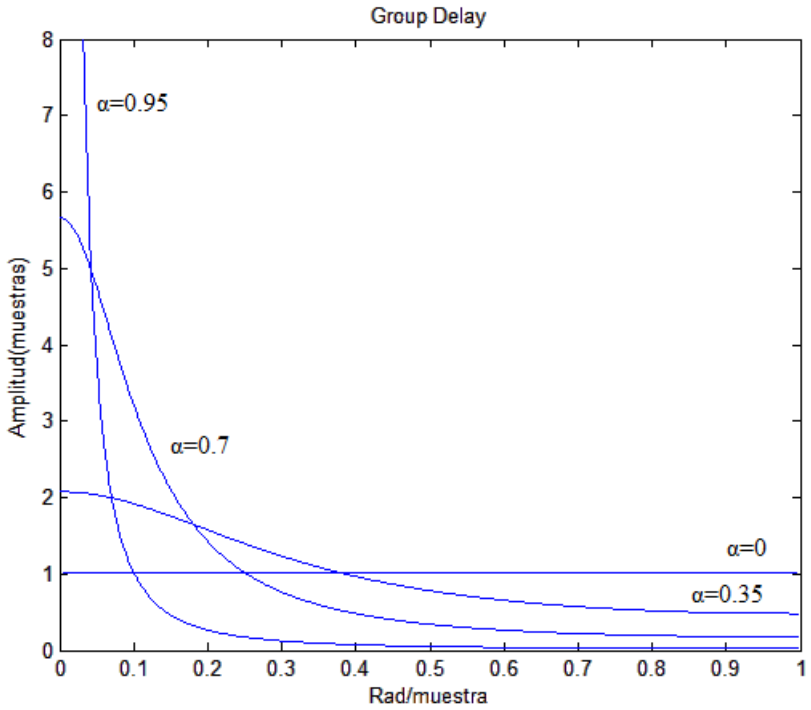

Figura 4.5: *Retardo de grupo del filtro pasa todas con diferentes valores de* α.

En la figura 4.6 se muestra el retardo de grupo del filtro pasa todas con una señal de entrada coseno con $10Hz$ de frecuencia. El retardo de grupo es de 2 muestras mientras que en la figura 4.7 con un valor mayor de α y la misma señal de entrada el retardo es de aproximadamente de 28 muestras.

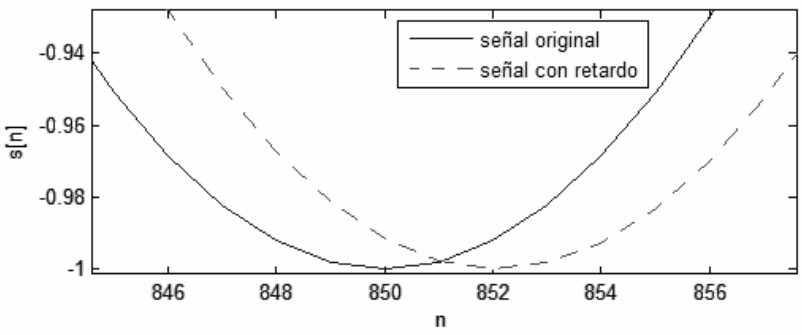

Figura 4.6: *Acercamiento de una señal coseno (linea continua) y la salida del filtro pasa todas (linea discontinua) con* $\alpha = 0.35$.

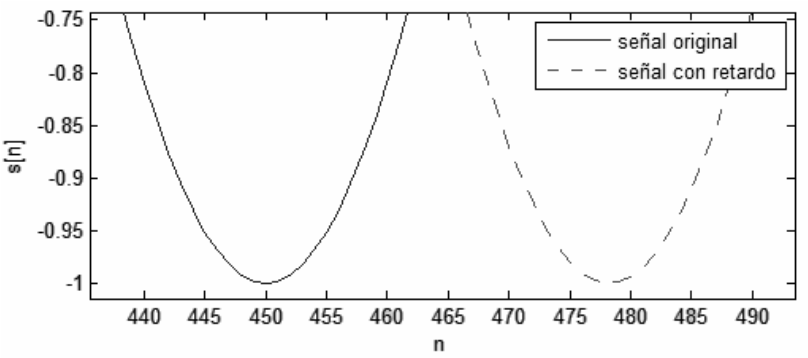

Figura 4.7: *Acercamiento de una señal coseno (linea continua) y la salida del filtro pasa todas (linea discontinua) con* $\alpha = 0.95$.

El mapeo de frecuencias del filtro se puede observar en las figuras 4.8, 4.9 y 4.10 para valores de $\alpha = 0$, $\alpha = 0.35$ y $\alpha = 0.9$.

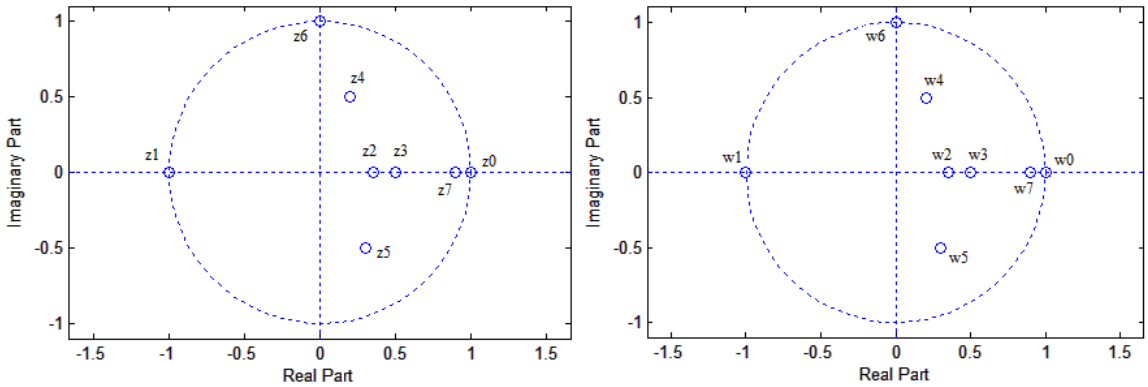

Figura 4.8: *Mapeo del filtro pasa todas con* $\alpha = 0$.

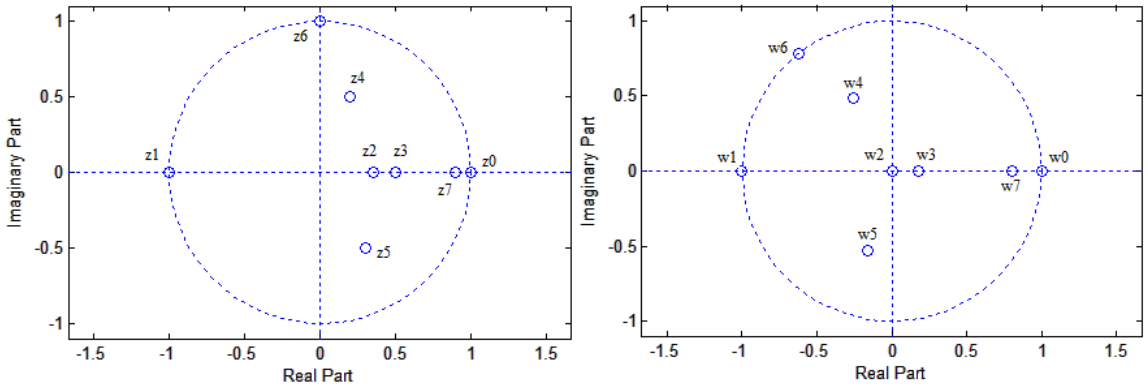

Figura 4.9: *Mapeo del filtro pasa todas con* $\alpha = 0.35$.

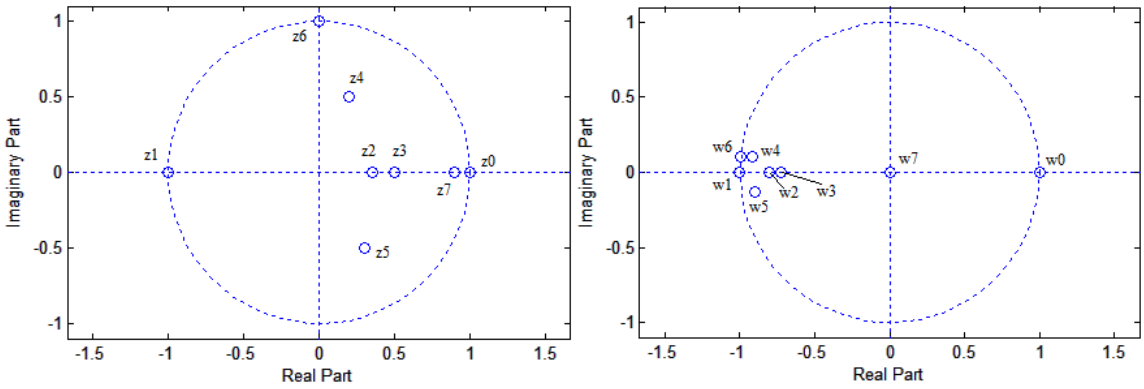

Figura 4.10: *Mapeo del filtro pasa todas con* $\alpha = 0.9$.

Al hacer $\alpha = 0$ el filtro se convierte en un simple retardo unitario y realiza el mapeo de frecuencias al mismo círculo unitario con la misma distribución. Al incrementar el valor de α se observa cómo las frecuencias se van agrupando en un área cada vez más reducida.

4.4. Cambio de Escala con el Filtro Pasa Todas

Si se hace un arreglo de filtros AP en cascada, se construye una cadena de filtros AP en la cual, al alimentar una señal con contenido frecuencial variado, el retardo de grupo no uniforme provocado por dicha cadena de filtros hace que los componentes de baja frecuencia avancen más lentamente y que los de alta frecuencia avancen más rápidamente que si la señal fuera alimentada en una cadena de retardos unitarios [6]. En la figura 4.11 se muestra un ejemplo de dos señales senoidades cuya escala de frecuencias ha sido cambiada (warped signal) en una cadena de filtros AP. Al estudiar los espectros de estas señales se comprueba inmediatamente que las frecuencias han cambiado.

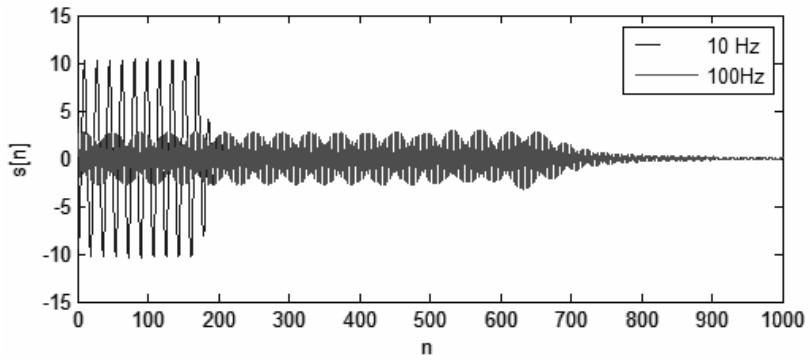

Figura 4.11: *Señales senoidales a la salida de una cadena de filtros AP con* $\alpha = 0.7$.

Existe una fórmula que determina el mapeo entre la escala natural de frecuencia y la escala transformada. ésta es tomada de la función de fase del filtro AP dada por

$$\tilde{f} = \beta(f) = \arctan \frac{\left(1-\alpha^2\right)\sin\left(f\right)}{\left(1+\alpha^2\right)\cos\left(f\right) - 2\alpha}. \tag{4.4.1}$$

En la figura 4.11 también se muestra cómo el cambio de escala cambia la estructura temporal de la señal original, de tal manera que los componentes de baja frecuencia son más cortos que la señal original, mientras

que los componentes de alta frecuencia pueden ser más largos. El retardo de grupo del filtro AP controla los cambios en la longitud de la señal.

El proceso de cambio de escala de una señal se realiza por [6]

$$S(z) = \sum_{k=0}^{N-1} a_k \tilde{z}^k. \tag{4.4.2}$$

Dado que el efecto del cambio de escala es una operación de desplazamientos variantes, es necesario procesar por tramas finitas de corta duración. Se deben tomar en cuenta las extrañas características de la señal escalada para procesarla adecuadamente. Por tales razones existe una segunda manera de realizar el cambio de escala de frecuencia, la cual consiste en aplicar el cambio a una función de transferencia o a una secuencia de coeficientes.

4.5. Escalas de Frecuencia *Bark* y *Mel*

La escala Bark de frecuencia es una escala psicoacústica propuesta por Eberhard Zwicker en 1961. La escala es de 1 a 24 Barks que corresponden a las primeras 24 bandas críticas del sistema auditivo. El término *banda crítica* fue introducido por Harvey Fletcher en la década de 1940 al realizar ejercicios sobre enmascaramiento [3]. El enmascaramiento consiste en reproducir sonidos de diferentes frecuencias simultáneamente y que uno de ellos evite la percepción de otros. Otro tipo de enmascaramiento ocurre cuando los sonidos tienen desplazamientos temporales entre ellos.

El enmascaramiento en frecuencia entre tonos está en función de la separación de sus frecuencias. Por ejemplo, si un tono enmascarador se fija a $1200Hz$ y $80dB$, un segundo tono de $800Hz$ puede ser escuchado desde una amplitud de $12dB$. Sin embargo, cuando el segundo tono está en el intervalo de $1200 \pm 100Hz$, el umbral de enmascaramiento es de $50dB$, este efecto ocurre también en altas frecuencias. Si las frecuencias de ambos tonos están muy alejadas entre sí, el umbral de enmascaramiento puede igualar al umbral en silencio. También se puede utilizar ruido de banda angosta para enmascarar un tono, donde el ancho de banda que contribuye al enmascaramiento se denomina banda crítica. Sólo el ruido dentro de una banda crítica contribuye al enmascaramiento, sin embargo, el umbral decae conforme la frecuencia central del ruido o del tono se aleja de las frecuencias centrales de las bandas críticas [3].

Tabla 4.1: *Bandas críticas*

Banda	Frec. central (Hz)	Ancho de banda (Hz)	Límite inferior (Hz)	Límite superior (Hz)
1	50	100	0	100
2	150	100	100	200
3	250	100	200	300
4	350	100	300	400
5	450	110	400	510
6	570	120	510	630
7	700	140	630	770
8	840	150	770	920
9	1000	160	920	1080
10	1170	190	1080	1270
11	1370	210	1270	1480
12	1600	240	1480	1720
13	1850	280	1720	2000
14	2150	320	2000	2320
15	2500	380	2320	2700
16	2900	450	2700	3150
17	3400	550	3150	3700
18	4000	700	3700	4400
19	4800	900	4400	5300
20	5800	1100	5300	6400
21	7000	1300	6400	7700
22	8500	1800	7700	9500
23	10500	2500	9500	12000
24	13500	3500	12000	15500

Estos anchos de banda deben ser interpretados como el muestreo de la variación continua en la respuesta en frecuencia del oido a una señal sinusoidal o proceso estocástico. La cóclea actúa como si estuviera contruida de filtros traslapados con anchos de banda iguales a las bandas críticas. Como se observa en la tabla el ancho de cada banda varía desde los $100Hz$ en bajas frecuencias hasta $3500Hz$ en altas frecuencias, aproximadamente un tercio de una octava.

El valor de α que mejor aproxima la escala Bark utilizando el filtro AP se calcula con la siguiente ecuación en función de la frecuencia de muestreo fs [1]:

$$\alpha(fs) = 1.0674 \left[\frac{\pi}{2} \arctan\left(0.06583 fs\right)\right]^{\frac{1}{2}} - 0.1916. \tag{4.5.1}$$

Para una frecuencia de muestreo $fs = 10kHz$, $\alpha = 0.4582$.

Para obtener la equivalencia de de frecuencia en hertz en Barks se usa la fórmula

$$Bark = 13 \arctan\left(0.00076 f\right) + 3.5 \arctan\left(\left(f/7500\right)^2\right). \tag{4.5.2}$$

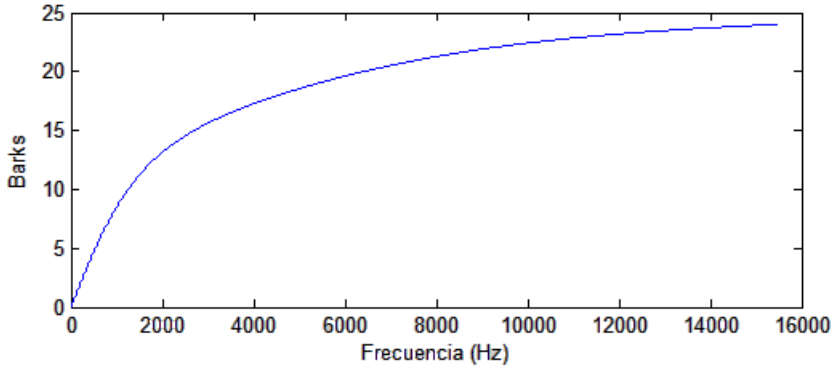

Figura 4.12: *Relación de Hertz con Barks.*

La escala mel fue propuesta por Stevens, Volknam y Newmann en 1937, es una escala musical basada en la percepción de tonos. Se define que un sonido de $1000Hz$ es también un sonido de $1000mels$ como punto de referencia. Si un grupo de observadores escuchan un sonido de $1000Hz$, y al incrementar la frecuencia consideran que ésta ha sido duplicada, entonces ese sonido será de $2000mels$ dado que la frecuencia real de ese sonido sería un poco menor a los $4000Hz$. Si se realizara otro experimento con la misma señal pero disminuyendo la frecuencia, cuando la mitad de la frecuencia sea percibida, se dirá que la señal es de 500 mels. En resumen, los observadores juzgan tonos espaciados exponencialmente como tonos equiespaciados.

El valor de $\alpha = 0.35$ es aquel que aproxima mejor la escala mel aplicando el filtro AP [17].

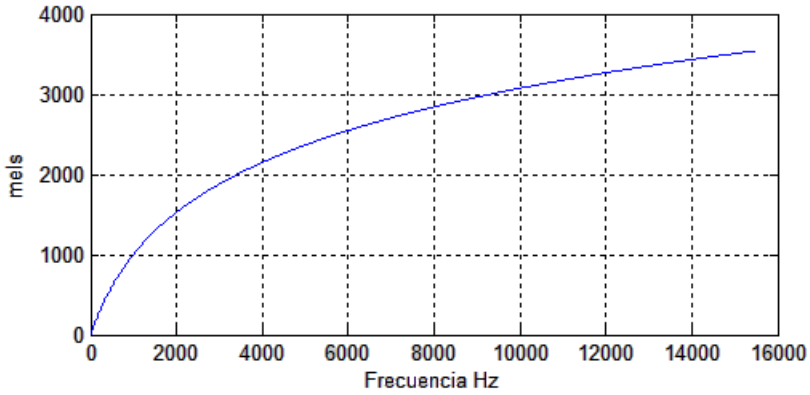

Figura 4.13: *Relación de la frecuencia en hertz con la escala mel.*

Capítulo 5

Diseño del Sistema de Codificación y Decodificación

El sistema de análisis por síntesis diseñado en este trabajo, está basado en el modelo de Satoshi [16] en el que se extraen los coeficientes mel cepstral y se utiliza a éstos como información de transmisión o almacenamiento, o, como representación de la señal comprimida. Tales coeficientes son utilizados para la generación de la voz sintética mediante un filtro de síntesis trasladado (warped filter) a la escala mel de frecuencia. El modelo contiene elementos muy básicos y primitivos para la recuperación de la voz tales como la clasificación de tramas según su sonoridad (tramas sonoras y no sonoras), el uso de la altura tonal y señales de excitación generadas con ruido blanco y trenes de impulsos. Además presenta limitaciones relacionadas con la señal de entrada dado que sólamente puede recibir señales de voz; cualquier otro tipo de señal aunque esté mezclada con voz no puede ser recuperada fielmente.

El modelo mencionado ha sido modificado en gran parte de sus elementos; sin embargo conserva los elementos principales: el proceso de extracción de los coeficientes cepstral y el filtro de síntesis. Se incorporaron nuevos elementos que mejoran la calidad y la naturalidad de la señal sintética; uno de ellos es el filtro de síntesis inverso para obtener datos útiles en la generación de la señal de excitación; otro elemento es el análisis wavelet para la generación de la señal de excitación. No se recurre más al cálculo de la altura tonal, a la clasificación de tramas, ni se trabaja sólo bajo la escala mel sino que se compara el uso de la escala Bark. El filtro de síntesis es llamado filtro de aproximación de espectro logarítmico en escala mel (MLSA, Mel Log Spectrum Approximation Filter) y está definido por [16]:

$$H(z) = exp \sum_{m=0}^{M} \tilde{c}(m)\tilde{z}^{-m} \tag{5.0.1}$$

donde

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, |z| < 1. \tag{5.0.2}$$

5.1. Extracción de Coeficientes Cepstral en Escalas Mel y Bark

Los coeficientes del filtro de síntesis se obtienen del cepstrum en escala mel, obtenido de la transformada inversa de Fourier IDFT del espectro logarítmico en escala mel. El procedimiento que se realiza para obtener los coeficientes del filtro es el mismo explicado en el Capítulo 2, con la variante de que se trabaja con una escala no lineal de frecuencia. La parte del cepstrum más cercana al origen corresponde a la función de transferencia del tracto vocal y es capaz de aproximar la envolvente espectral de la señal. Al momento de realizar el filtrado en el dominio de la quefrencia para extraer los primeros M elementos del cepstrum, éste

es sometido a una cadena de filtros pasa todas para realizar el cambio de escala de frecuencia lineal a escala mel o Bark, según el valor del coeficiente pasa todas [6].

5.2. Elementos del Sistema de Análisis o Codificación

Antes de realizar cualquier operación con la señal de entrada es importante recurrir a algunos métodos tradicionales de pre-procesamiento que preparan a la señal para ser procesada eficientemente:

Segmentación en tramas. Procesar la señal por partes es fundamental en procesos estocásticos no estacionarios. Las señales de voz pertenecen a este tipo de procesos puesto que su distribución de probabilidad es variante con el tiempo. Sin embargo, tales variaciones pueden considerarse inexistentes en un intervalo corto de tiempo (del orden de milisegundos), por lo que al tomar un fragmento de señal de poca duración, se observa que su distribución de probabilidad es estacionaria y por lo tanto, se denomina proceso cuasi-estacionario, y este proceso puede ser sometido a transformaciones y operaciones definidas para procesos estacionarios. Para dividir la señal en tramas ésta es multiplicada por una ventana de Hanning, cuyas características espectrales evitan interferencias o alteraciones en la frecuencia original, provocada por el truncamiento del segmento de voz del resto de la señal; lo que sería equivalente a multiplicar por una ventana rectangular.

Filtro de Pre-énfasis. Este filtro incrementa la energía de las altas frecuencias con el fin de no perder información de detalles al momento de extraer información de la señal. Al final del proceso, se aplica el filtro de de-énfasis cuyo efecto es el inverso y recupera las características originales del espectro de la señal. Al filtro de pre-énfasis lo define la siguiente ecuación en diferencias [5]:

$$y[n] = x[n] - ax[n-1] \tag{5.2.1}$$

donde el valor de $a = 0.9$ es usualmente seleccionado.

5.2.1. Bloque de Extracción de Coeficientes Cepstral

Como se ha explicado en el Capítulo 2, se toma una trama de voz pre-enfatizada, se calcula la FFT y a su magnitud se le calcula el logaritmo para enseguida calcular la IFFT. Una vez obtenido el cepstrum, se extrae el número solicitado de parámetros mientras es procesado en la cadena de filtros pasa todas para obtener parámetros en escala mel o Bark según el valor de la constante pasa todas.

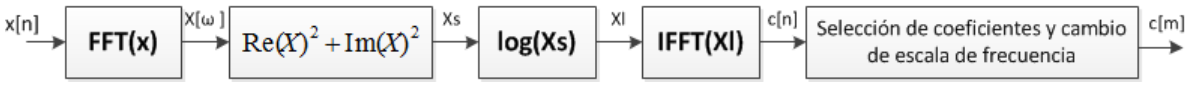

Figura 5.1: *Algortimo para calcular los coeficientes cepstral.*

El algoritmo en pseudocódigo para realizar el cambio de escala se muestra a continuación [16]:

1. Entra: $c[n]$. ($c[n]$ es el cepstrum)
2. De $i \leftarrow -N$ a 0 (N es la longitud del cepstrum)
3. $g[0] \leftarrow c[-i] + a * (d[0] \leftarrow g[0])$
4. $g[1] \leftarrow (1 - a * a) * d[0] + a * (d[1] \leftarrow g[1])$
5. De $j \leftarrow 2$ a M (M es el número de coeficientes seleccionado)
6. $g[j] \leftarrow d[j-1] + a * ((d[j] \leftarrow g[j]) - g[j-1])$

7. $c \leftarrow g$

8. Regresa: $c[m]$

El valor de M define el orden del filtro de síntesis y la cantidad de parámetros cepstrum que se enviarán al sistema decodificador.

5.2.2. Bloque de Detección de Tramas no Sonoras

Dada la pérdida de información de detalles después de la compresión, tanto la naturalidad de la voz como la inteligibilidad de sonidos con contenido ruidoso o con consonantes fricativas, pueden ser mejoradas sumando ruido de diferentes anchos de banda y amplitudes a la señal de excitación. Para mejorar la naturalidad de la voz, se suman a la señal de excitación ruidos de diferentes anchos de banda de muy bajas amplitudes y constantes en todo el proceso y, para la recuperación de consonantes fricativas se suma ruido blanco de una amplitud mayor a los otros ruidos, sumados a la señal de excitación sólo en las tramas detectadas como no sonoras.

Un bit es enviado al decodificador para indicar la presencia de una trama con baja sonoridad detectada midiendo la cantidad de cruces por cero y la energía de las tramas de las señal original.

5.2.3. Bloque de Generación de la Señal de Análisis

Para obtener la señal de análisis se requiere obtener el filtro MLSA inverso el cual queda definido con la siguiente ecuación:

$$\frac{1}{H(Z)} = exp \sum_{m=0}^{M} -\tilde{c}(m)\tilde{z}^{-m} \tag{5.2.2}$$

Este filtro recibe la misma señal de entrada al bloque de extracción de coeficientes cepstral y es modelado con los coeficientes cepstral. La salida de este filtro contiene información útil para generar la señal de excitación del filtro de síntesis. Además de los parámetros cepstral, es necesario enviar parámetros adicionales extraidos de la señal de análisis mediante la DWT los cuales contienen información suficiente para aproximar la señal de análisis en el decodificador aplicando la IDWT.

5.2.4. Bloque de Compresión por DWT de la Señal de Análisis

Al calcular la transformada wavelet discreta de la señal de análisis, se ha observado que la mayoría de los coeficientes tienen magnitudes muy pequeñas; muy cercanas a cero. Por lo tanto, comprimir implica truncar aquellos coeficientes que estén por de bajo de un umbral.

Los componentes de baja frecuencia son la parte más importante de la voz humana. Cuando los componentes de alta frecuencia son removidos, la señal de voz conserva su inteligibilidad pese a que suena un poco diferente. Por esta razón sólamente los coeficientes de detalles son truncados. Al realizar experimentos se encuentra que al rededor del 90 % de los coeficientes wavelet tienen magnitud despreciable y su truncamiento a cero provoca, en la reconstrucción, diferencias poco perceptibles en la señal.

Para alcanzar un grado mayor de compresión, todos los coeficientes de detalles pueden ser truncados a cero dejando los coeficientes de aproximación del último nivel del árbol de descomposición. Si se desea incrementar la calidad de la señal recuperada, se debe disminuir el grado de compresión dejando cierta cantidad de coeficientes de detalles, ya sea definiendo un umbral o seleccionando un número fijo de valores.

Los coeficienes que no fueron truncados son almacendos en un vector, y en un segundo vector se almacena la información de la posición inicial de una secuencia de ceros y el número de ceros. Finalmente los coeficientes de los vectores son cuantizados y transmitidos al decodificador.

Ejemplo de la codificación de los coeficientes wavelet: Sea CX el vector de la DWT con coeficientes truncados $CX = [1, 0, 0, 0, 0, 0, 5, 2, 2, 0, 0, 0, 9, 0, 0, 3]$. El vector Co de coeficientes diferentes de cero es $Co = [1, 5, 2, 2, 9, 3]$ y el vector de posición de ceros es $Zpos = [2, 5, 10, 3, 14, 2]$.

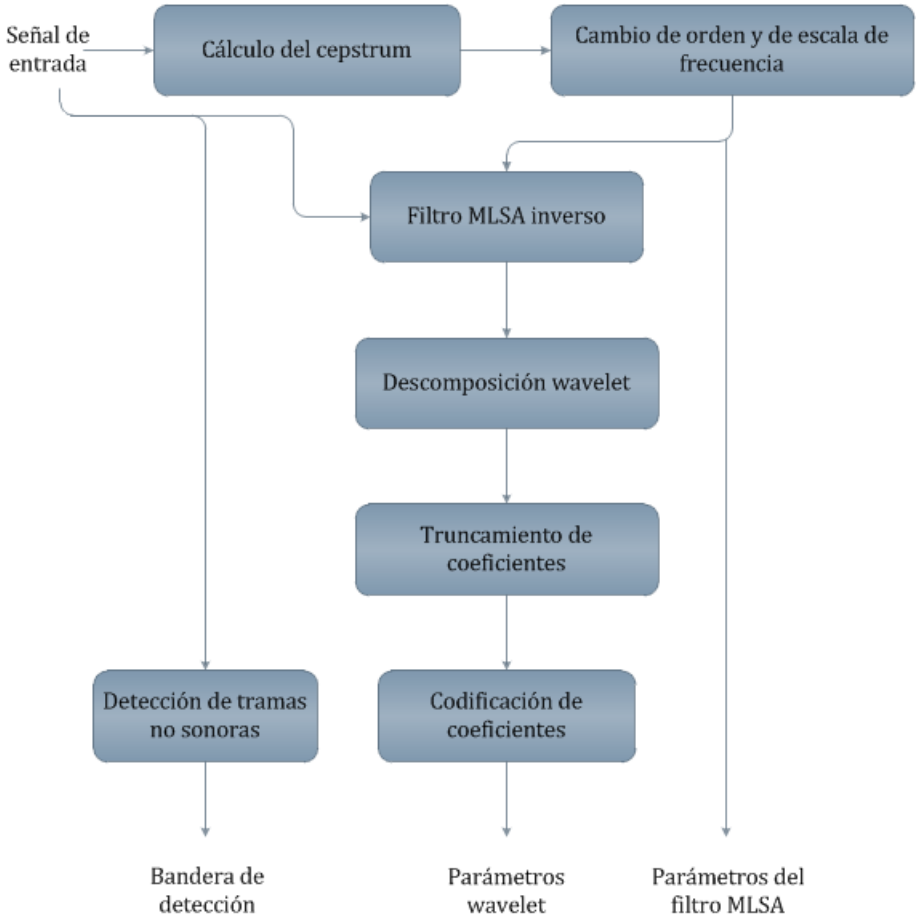

Figura 5.2: *Elementos del sistema de codificación.*

5.3. Elementos del Sistema de Síntesis o Decodificación

El sistema de decodificación recibe los coeficientes cepstral y wavelet para reconstruir la señal. Estos coeficientes son cuantizados antes de ser transmitidos desde el codificador. Dicho proceso se explica más adelante.

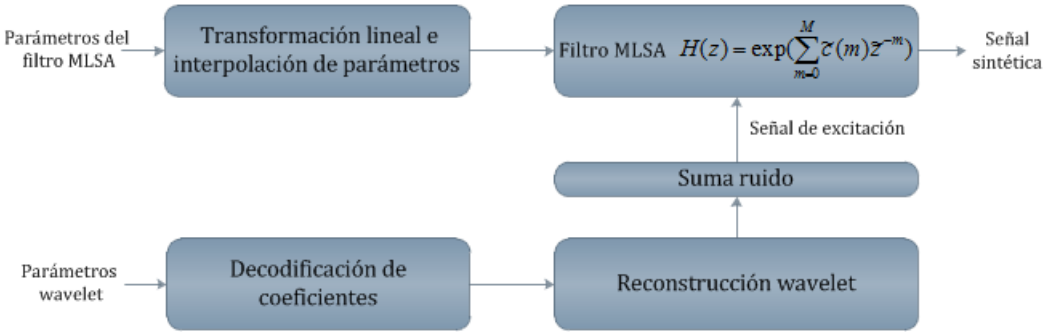

Figura 5.3: *Elementos del sistema decodificador.*

5.3.1. Bloque de Transformación Lineal e Interpolación de Parámetros

El módulo de interpolación de parámetros obtiene los parámetros cepstral de la trama actual y de la trama siguiente con el propósito de llevar a cabo, durante las iteraciones, una interpolación de dichos parámetros para que el cambio entre tales sea suave y por consiguiente lo sea la señal sintética.

La siguiente transformación lineal convierte los coeficientes mel/Bark-cepstral en coeficientes MLSA [16]:

$b_\alpha(M+1) = \alpha c_\alpha(M)$
$b_\alpha(m) = c_\alpha(m) + \alpha(c_\alpha(m-1) - b_\alpha(m+1))$, (m=M, M-1, ..., 3, 2)
$b_\alpha(1) = (c_\alpha(1) - \alpha b_\alpha(2))/(1-\alpha^2)$
$b_\alpha(0) = c_\alpha(0) - \alpha b_\alpha(1)$.

5.3.2. Bloque de Decodificación de Coeficientes Wavelet y Generación de la Señal de Excitación

Los vectores de coeficientes wavelet y de posiciones de ceros entran a este módulo y se reconstruye la señal que fue el resultado de la DWT y del truncamiento a cero. Esta señal es sometida a la IDWT para recuperar la señal de análisis y opere como señal de excitación del filtro MLSA. El proceso que se sigue está descrito en el Capítulo 3, en el cual en cada nivel de descomposición se sobremuestrea la señal mediante el rellenado con ceros y se convoluciona con los filtros de reconstrucción cuyos coeficientes corresponden a la función wavelet db2.

5.3.3. Bloque de Suma de Ruido a la Señal de Excitación

Señales de ruido blanco son divididas en cuatro bandas de frecuencia. Los límites de las bandas en Hertz son [500 1000 2500 3500 5000]. El promedio de las amplitudes de cada banda de ruido fue estimado con las amplitudes de información en señales originales que corresponden a la información perdida en las señales sintéticas. Para lograrlo, la señal original fue dividida en las cuatro bandas de frecuencia definidas y se identificó la información que no está presente en la señal sintética. Se midieron las amplitudes promedio de la información que se pierde para ser utilizadas como la ganancia de cada banda de ruido inicialmente normalizado. Una vez que cada banda de ruido ha sido multiplicada por la ganancia correspondiente, es sumada a la señal de excitación.

El cáculo de las ganancias no se realiza en el sistema, sino que han sido calculadas experimentalmente para ser utlilizadas como parámetros constantes en el decodificador y con ello no incrementar la tasa de bits. Este procedimiento podría ser realizado en el decodificador para enviar los valores de ganancia para cada trama, sin embargo esto incrementa bastante la tasa de transferencia de bits, además se ha obtenido un buen desempeño utilizando los parámetros constantes para toda la señal.

Adicional a la suma de bandas de ruido, se suma ruido blanco sólo a las tramas que fueron clasificadas como de baja sonoridad cuando el decodificador recibe el bit de detección con valor igual a 1. La amplitud de este ruido blanco fue estimada con el promedio de amplitudes de este tipo de tramas en señales originales. Este bloque podría ser omitido del sistema de compresión si se decidiera no trabajar con un alto nivel de compresión. Desde que un grado de compresión alto implica mayor pérdida de detalles, este bloque compensa dicha pérdida.

5.3.4. Bloque de Síntesis con el Filtro MLSA

La función de transferencia del filtro MLSA está dada por una función racional de la función de transferencia de un filtro pasa todas de primer orden. Este filtro tiene baja sensibilidad espectral a variaciones en los coeficientes como ocurre en la cuantización [16].

La función de transferencia del filtro MLSA es la aproximación de la exponencial de la función de transferencia de un filtro base.

La forma ideal de la función de transferencia es [16]

$$H_\alpha^0(z) = e^{F_\alpha(z)} \tag{5.3.1}$$

donde $F_\alpha(z)$ es la función de transferencia del filtro base. Si éste es estable, el filtro MLSA también lo es, además de ser de fase mínima puesto que su función de transferencia tiene una cantidad finita de valores diferentes de cero dentro del círculo unitario en el plano z.

Dada la función de transferencia del filtro base

$$F_\alpha(\tilde{z}) = \sum_{m=0}^{M} c_\alpha(m)\tilde{z}^{-m} \tag{5.3.2}$$

el logaritmo de la magnitud de la respuesta de un filtro MLSA ideal en la escala de frecuencia mel esta dada por

$$ln|H_\alpha^0(e^{j\tilde{\Omega}})| = \sum_{m=0}^{M} c_\alpha(m)cos(m\tilde{\Omega}) \tag{5.3.3}$$

Al asignar a $c(m)$ como los parámetros mel cepstral, el logaritmo de la magnitud de la respuesta será idéntico a la envolvente del espectro logarítmico en la escala mel. La envolvente del espectro logarítmico puede representarse por medio de un polinomio trigonométrico de orden M de la forma

$$G_\alpha(\tilde{f}) = \sum_{m=0}^{M} c_\alpha(m)\cos(m\tilde{f}) \tag{5.3.4}$$

donde $\tilde{f}$ se obtiene de la función de fase del filtro AP; la ecuación (4.4.1) [6].

En contraste, la forma ideal del filtro MLSA no es realizable debido a su forma exponencial. Una función de transferencia de forma exponencial puede ser aproximada por una función racional generada mediante la aproximación Padé.

La aproximación Padé $R_L(w)$ de orden (L, L) para una función exponencial $e^{(w)}$ está dada por

$$R_L(w) = P_L(w)/P_L(-w) \tag{5.3.5}$$

$$P_L(w) = 1 + P_{L,1}w(1 + P_{L,2}w(...(1 + P_{L,L}w))...) \tag{5.3.6}$$

La función de transferencia $F_\alpha(\tilde{z})$ del filtro base se reescribe como [16]

$$F_\alpha(\tilde{z}) = F(z) = b_\alpha(0) + z^{-1}\sum_{m=1}^{M+1} b_\alpha(m)\tilde{z}^{-(m-1)} \tag{5.3.7}$$

El parámetro $b_\alpha(m)$ se obtiene a través de ecuaciones recursivas antes definidas. Sea

$F_\alpha^{(0)}(\tilde{z}) = b_\alpha(0)$
$F_\alpha^{(1)}(\tilde{z}) = z^{-1}b_\alpha(1)$
$F_\alpha^{(2)}(\tilde{z}) = z^{-1}(b_\alpha(2)\tilde{z}^{-1} + b_\alpha(3)\tilde{z}^{-2})$
$F_\alpha^{(3)}(\tilde{z}) = z^{-1}(b_\alpha(4)\tilde{z}^{-3} + ... + b_\alpha(7)\tilde{z}^{-6})$
$F_\alpha^{(4)}(\tilde{z}) = z^{-1}(b_\alpha(8)\tilde{z}^{-7} + ... + b_\alpha(M+1)\tilde{z}^{-M})$
y

$$H_\alpha(\tilde{z}) = e^{b_\alpha(0)}\prod_{k=1}^{M+1} R_L(F_\alpha^k(\tilde{z})) \tag{5.3.8}$$

entonces
$(F_\alpha^k(\tilde{z}))_{z^{-1}=0} = (F_\alpha^k(z))_{z^{-1}=0,\,\tilde{z}^{-1}=-\alpha} = 0$ [16].

Con las modificaciones anteriores, el filtro MLSA con función de transferencia $H_\alpha(\tilde{z})$, no presenta bucles libres de retardos (delay-free loops) al transcribirlo en código de programación. El filtro MLSA, sin dichas modificaciones, presenta bucles sin retardos, lo que lo convierte en un algoritmo no *secuencialmente computable*. Para evitar un problema de esta naturaleza, es indispensable que en cada ciclo en el código del filtro, exista por lo menos un elemento de retardo [24].

En el proceso de síntesis el filtro recibe la señal de excitación y los parámetros cepstral en escala mel o Bark de cada trama. Estos últimos son escalados a hertz (unwarped) en el filtro sustituyendo cada retardo unitario por el filtro pasa todas, con la misma magnitud de α pero con signo negativo, mientras el filtro sintetiza. Este proceso está descrito en el Capítulo 2.

El cambio en el filtro base en la ecuación (5.3.7) se realiza para extraer la ganancia K afuera de $F(z)$ y no sea unitaria, además de hacer *realizable* al filtro. Como antecedente a (5.3.7), se reescribe el filtro MLSA en la forma [18]

$$H(z) = \exp \sum_{m=0}^{M} b(m)\phi_m(z) = K \cdot D(z) \tag{5.3.9}$$

donde

$$K = \exp b(0) \tag{5.3.10}$$

$$D(z) = \exp \sum_{m=1}^{M} b(m)\phi_m(z) \tag{5.3.11}$$

y

$$c_\alpha(m) = \begin{cases} b(m), & m = M \\ b(m) + \alpha b(m+1), & 0 \leq m < M \end{cases} \tag{5.3.12}$$

$$\phi_m(z) = \begin{cases} 1 & m = 0 \\ \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}} \tilde{z}^{-(m-1)} & m \geq 1 \end{cases} \tag{5.3.13}$$

El filtro MLSA provee una aproximación muy exacta a la envolvente espectral del espectro logarítmico en escala mel o Bark.

5.4. Cuantización de los Parámetros del Codificador

Los parámetros cepstral y wavelet son cuantizados de acuerdo a sus características. Al observar las magnitudes de una gran cantidad de coeficientes cepstral, se concluyó que la magnitud máxima del primer coeficiente cepstral es 8 (para m=1) y la máxima magnitud para el resto de los coeficientes es 1 (m=2,..M). El primer parámetro es truncado a un entero de 3 bits. El grupo restante de parámetros es normalizado a la magnitud unitaria (trasladados al intervalo [0, 1]). Esto permite el uso de un solo libro de códigos en el decodificador. La cantidad de bits seleccionada para cuantizar este grupo de parámetros depende del grado de compresión deseado. Se debe tomar en cuenta un bit adicional para representar el signo. Para recuperar las magnitudes originales de estos valores es necesario transmitir el valor máximo del grupo, el cual es el mismo por el cual el conjunto fue dividido para normalizar sus magnitudes. No más de 8 bits son necesarios para representar a dicho valor.

También, al observar una gran cantidad de coeficientes wavelet, se encontró que la diferencia entre el valor máximo y el mínimo de cada grupo de parámetros wavelet no es mayor a 1. El mismo procedimiento de normalización de magnitudes es aplicado a estos valores, los cuales son mapeados al intervalo [0, 1]. En este caso la magnitud máxima de cada grupo no es transmitida, puesto que se demostró experimentalmente que no hay distorsión en la señal sintética si no se recupera la magnitud original de los coeficientes, con lo que se evita el incremento de la tasa de transferencia de bits. Entonces, la cantidad de bits elegida para cuantizar

ambos tipos de parámetros es variable de acuerdo al nivel de compresión deseado y a la calidad de la señal sintética deseada.

Tabla 5.1: *Distribución de bits de los datos que transmite el codificador.*

Parámetro	Resolución (bits/parámetro)
Cepstral (con signo)	4
Primer cepstral	3
Magnitud máxima	8
Wavelet	5
Bit de detección	1

A continuación se muestra un ejemplo de la cuantización de los parámetros del sistema.

Sea C un vector de coeficientes cepstral, E un vector de coeficientes wavelet y A la amplitud máxima del vector C.

C={-4.7661 0.0309 0.1169 0.3583 0.2295 0.3428 0.2276 0.2027 0.2192 0.3302 0.1977 0.2667 0.2071 0.1548 0.2139 0.1550 0.1803 0.2435 0.0787 0.0923 0.0031 0.0327 0.0365 0.0726}.

Los valores normalizados y cuantizados de C con 4 bits quedan como se muestra en Cq.

Cq={-5* 0.0625 0.3125 0.9375 0.6875 0.9375 0.6875 0.5625 0.5625 0.9375 0.5625 0.6875 0.5625 0.4375 0.5625 0.4375 0.5625 0.6875 0.1875 0.3125 0.0625 0.0625 0.0625 0.1875 0.3125}.

En el vector Cq los coeficientes están normalizados y cuantizados con 4 bits excepto el primer coeficiente, el cual fue truncado a entero y cuantizado con 3 bits sin signo puesto que siempre es negativo.

Al dividir C por su magnitud máxima sin tomar en cuenta al primer coeficiente se obtiene $A = 0.3583$, al cuantizarlo utlizando el mismo intervalo [01] pero con 8 bits queda $Aq = 0.3223$.

El vector Cqr con las magnitudes recuperadas se obtiene multiplicando Cq por Aq.

Cqr={-5 0.0223 0.1117 0.3351 0.2457 0.3351 0.2457 0.2010 0.2010 0.3351 0.2010 0.2457 0.2010 0.1564 0.2010 0.1564 0.2010 0.2457 0.0670 0.1117 0.0223 0.0223 0.0223 0.0670 0.1117}.

Los coeficientes E son

E={-0.0139 -0.0255 0.0032 -0.0121 -0.0350 -0.0641 -0.1206 -0.0599 -0.0177 -0.0544 0.0103 0.0572 -0.0353 0.0165 0.0338 -0.0315 -0.0076 -0.0277 -0.1345 -0.0461 -0.0388 -0.1132 -0.0115 0.0183 -0.0438 0.0008 0.0070 -0.0527 0.0662 0.0361 -0.0638 -0.0469 0.0004 -0.0873 0.0060 -0.0232}

El vector E normalizado queda

En={-0.1036 -0.1896 0.0237 -0.0901 -0.2604 -0.4765 -0.8970 -0.4455 -0.1317 -0.4046 0.0766 0.4253 -0.2627 0.1226 0.2515 -0.2343 -0.0568 -0.2060 -1.0000 -0.3425 -0.2884 -0.8421 -0.0859 0.1362 -0.3253 0.0062 0.0522 -0.3919 0.4922 0.2682 -0.4747 -0.3485 0.0033 -0.6494 0.0449 -0.1723}

y al cuantizarlos con 5 bits

Eq={-0.1094 -0.2031 0.0156 -0.0781 -0.2656 -0.4844 -0.8906 -0.4531 -0.1406 -0.3906 0.0781 0.4219 -0.2656 0.1094 0.2656 -0.2344 -0.0469 -0.2031 -0.9844 -0.3281 -0.2969 -0.8281 -0.0781 0.1406 -0.3281 0.0156 0.0469

-0.3906 0.4844 0.2656 -0.4844 -0.3594 0.0156 -0.6406 0.0469 -0.1719}.

El error de cuatización de los coeficientes no genera distorsión espectral significativa dada su naturaleza, además esta asignación de bits no es perceptible en las señales sintéticas al compararlas con señales generadas con coeficientes sin cuantizar.

5.5. Calidad de la Señal Sintética y Tasa de Bits

Con el propósito de evaluar la calidad de las señales sintéticas se grabaron[1] oraciones de duraciones variadas (entre 2 y 10 segundos) para ser codificadas y decodificadas por el sistema y posteriormente analizadas por varios oyentes. Otras señales de voz extraídas de archivos mp3 también fueron incluidas en las señales evaludas.

Dados los parámetros:
Fs: frecuencia de muestreo.
T: duración de la trama.
M: orden del cepstrum.
b_c: bits por coeficiente cepstrum.
W: número de coeficientes wavelet.
b_w: bits por coeficiente wavelet.
b_M: bits por valor cepstral máximo.
b_F: bits por primer parámetro cepstral.
b_s: bit de detección de trama no sonora.
α: constante pasa todas.
La tasa total de bits B para este sistema se calcula con

$$B = \left[(M-1)\cdot b_c + W \cdot b_w + b_M + b_F + b_s\right]/T \tag{5.5.1}$$

Para $Fs = 10kHz$, $T = 25ms$, $M = 26$, $b_c = 4$, $W = 34$, $b_w = 5$, $b_M = 7$, $b_F = 3$ y $\alpha = 0.35$ (escala mel) la tasa de bits es $B = 10.9kbit/s$. La calidad de la voz es muy alta y la inteligibilidad es muy alta, el hablante es claramente reconocible y las voces se perciben con gran naturalidad.

Para $Fs = 10kHz$, $T = 25ms$, $M = 26$, $b_c = 4$, $W = 18$, $b_w = 5$, $b_M = 7$, $b_F = 3$ y $\alpha = 0.35$ (escala mel) la tasa de bits es $B = 7.8kbit/s$. La calidad de la voz es buena y la inteligibibilidad es muy alta. El hablante es claramente reconocible sin embargo la señal pierde un poco de naturalidad. Lo anterior se basa en una sola prueba de percepción que consiste en escuchar la señal y juzgarla. Esta evaluación es preliminar e informal, pero útil para decidir si la calidad de este codificador es suficiente para compararla con la de otros sistemas de codificación.

Una prueba MOS (Mean Opinion Score) fue realizada con 10 personas; éstas escucharon y evaluaron siete señales por separado. La prueba MOS tiene una escala de 1 a 5 donde 1 es la peor y 5 la mejor [23]. La calificación promedio de cada una se muestra en la tabla 5.2.

[1]Las grabaciones fueron realizadas en el Laboratorio de procesamiento de voz, utilizando micrófonos unidireccionales de propósito general y la tarjeta de sonido de una computadora.

Tabla 5.2: *Resultados de la prueba MOS.*

Señal	7.8 kbit/s (mel)	10.9 kbit/s (mel)	7.8 kbit/s (Bark)	10.9 kbit/s (Bark)
1 (masculina)	3.60	4.4	3.50	4.5
2 (femenina)	3.9	4.2	3.90	4.2
3 (masculina)	3.30	4.6	3.44	4.6
4 (masculina)	3.30	3.9	3.34	3.8
5 (femenina)	3.37	4	3.35	4
6 (femenina)	2.90	3.8	2.94	3.9
7 (femenina)	2.90	3.9	3.50	3.9
Promedio	3.32	4.11	3.42	4.12

La tabla 5.3 muestra las características principales de otros codificadores de voz y su calificación MOS.

Tabla 5.3: *Características de otros codificadores y calificación MOS.*

Número	Estándar	Tasa de bits (kbit/s)	Frecuencia de muestreo (kHz)	Tamaño de trama (ms)	MOS
G.711	ITU-T	64	8	muestra	4.1
G.721	ITU-T	32	8	muestra	
G.722	ITU-T	64	16	muestra	3.6
G.722.1	ITU-T	24/32	16	20	
G.723	ITU-T	24/40	8	muestra	
G.723.1	ITU-T	5.6/6.3	8	30	3.8-3.9
G.726	ITU-T	16/24/32/40	8	muestra	3.85
G.727	ITU-T	var.		muestra	
G.728	ITU-T	16	8	2.5	3.61
G.729	ITU-T	8	8	10	3.92
GSM 06.10	ETSI	13	8	22.5	3.7
LPC10	USA Government	2.4	8	22.5	
Speex		2-44	8,16,32		
iLBC		8	13.3	30	
DoD CELP	American Department of Defense (DoD) USA Government	4.8		30	
DVI	Interactive Multimedia Association (IMA)	32	variable	muestra	

En las tablas 5.4 y 5.5 se muestran los resultados de la evaluación MOS aplicada a cinco señales de voz diferentes. En las tablas se compara la calificación del codificador propuesto en esta tesis con la de otros codificadores de voz, tales como WMA, Speex y GSM-FR (Full rate, 13 kbps). La prueba MOS se realizó en el laboratorio de procesamiento de voz con diez personas.

Tabla 5.4: *Comparación de calificaciones MOS del codificador de la tesis con otros codificadores (8kbps).*

Señal	Códec propuesto (7.2kbps)	WMA (8kbps)	SPX (8kbps)
1	3.7	3.8	2.9
2	3.2	3.9	3.7
3	3.5	3.9	3.4
4	3.2	3.8	3.7
5	3.1	3.9	4.0
Promedio	3.34	3.86	3.54

Tabla 5.5: *Comparación de calificaciones MOS del codificador de la tesis con otros codificadores (11kbps).*

Señal	Códec propuesto (11kbps)	GSM-FR (13kbps)	SPX (11kbps)
1	3.4	3.1	3.1
2	4.0	3.0	4.0
3	3.9	3.5	3.9
4	3.4	3.2	3.7
5	3.5	3.2	3.9
Promedio	3.64	3.20	3.72

En la tabla 5.6 se compara las calificaciones de los codificadores de las tablas 5.4 y 5.5 con señales de voz mezcladas con música.

Tabla 5.6: *Comparación de calificaciones MOS del codificador de la tesis con otros codificadores (11kbps). Señales de voz mezcladas con música.*

Señal	Codec propuesto (11kbps)	GSM-FR (13kbps)	SPX (11kbps)
voz y música	3.2	3.7	3.6

La figura 5.4 muestra la comparación de espectros de una señal original con su reconstrucción.

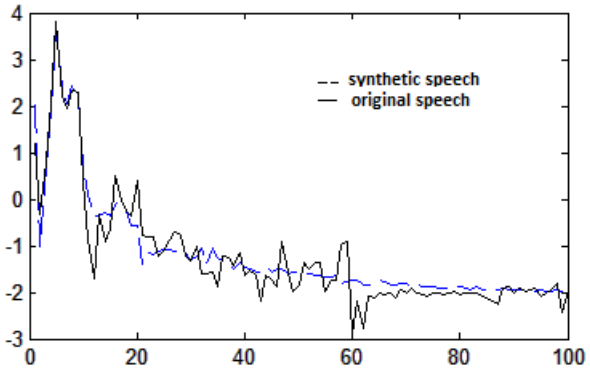

Figura 5.4: *Comparación de espectros.*

Sea $x[n]$ una señal original y $y[n]$ la versión sintética, la relación señal a ruido se define con

$$SNR = 10\log_{10}\left(\frac{\sum_n x[n]^2}{\sum_n (x[n]-y[n])^2}\right) \tag{5.5.2}$$

donde n en el intervalo de tiempo a medir.

El cálculo del SNR es significativo sólo en codificadores de forma de onda. Por otro lado, la mayoría de los codificadores de baja tasa de transferencia, como el presentado en este capítulo, no conserva la forma de onda original y por lo tanto el valor del SNR no es significativo en la evaluación de este codificador. Las técnicas de evaluación subjetiva de calidad fueron diseñadas para superar las limitaciones del uso del SNR.

Como un dato que representa el acercamiento de la forma de onda sintética a la original se estimó el SNR para diferentes segmentos de diferentes señales, el valor promedio de SNR en tramas de alta energía o sonoras es $SNR = 13dB$, mientras que el valor medio en tramas no sonoras es cercano a cero. Esta medida es muy sensible a distorsiones de fase y diferencias de fase entre ambas señales los cuales no son usualmente perceptibles.

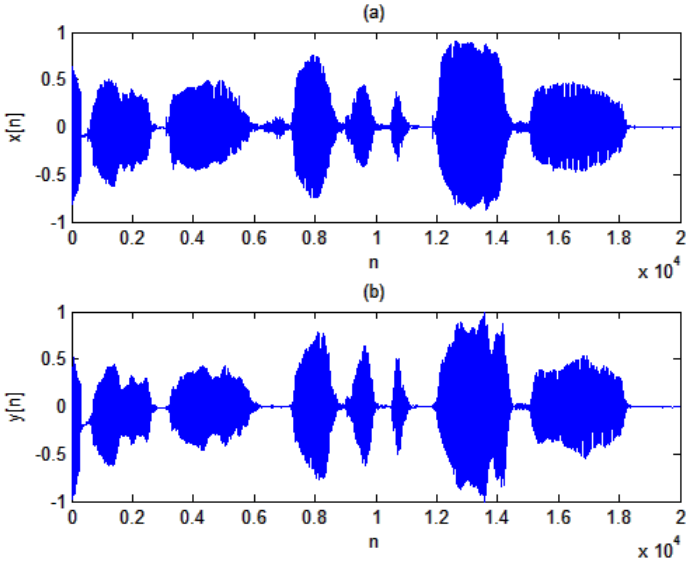

Figura 5.5: *(a) Señal original de voz masculina. (b) Versión sintética de (a).*

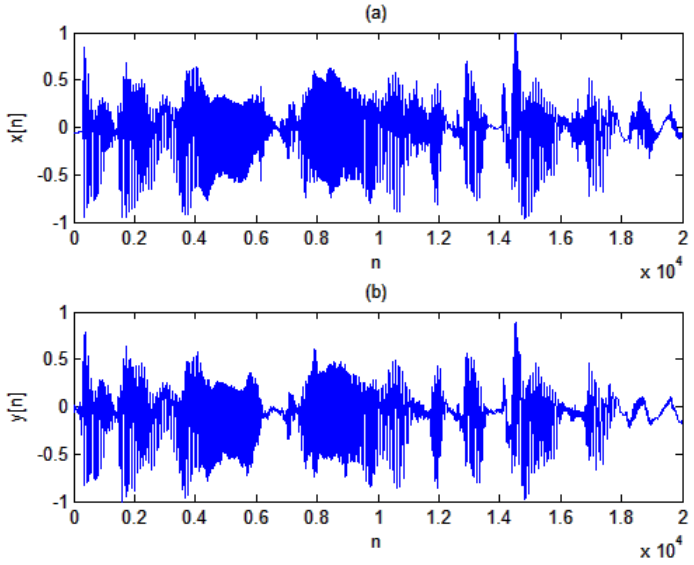

Figura 5.6: *(a) Señal original de voz masculina. (b) Versión sintética de (a).*

Las gráficas de las figuras 5.5 y 5.6 muestran señales originales de pura voz de dos segundos de duración y sus versiones sintetizadas con el sistema de codificación a una tasa de bits de $10.9 kbit/s$. La forma de onda de las señales recuperadas es muy parecida a la de las señales originales.

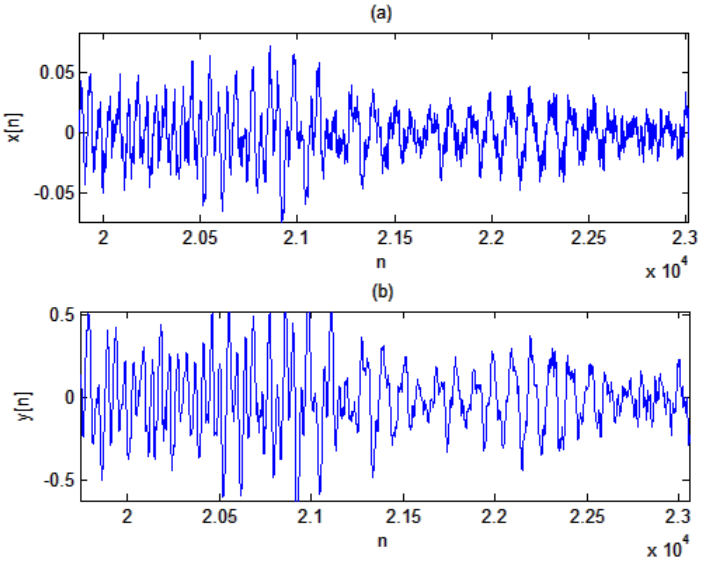

Figura 5.7: *(a) Señal original de voz femenina con música de fondo. (b) Versión sintética de (a).*

La figura 5.7 muestra un segmento de una señal de audio que contiene voz y música, la forma de onda de la señal recuperada es muy similar a la original.

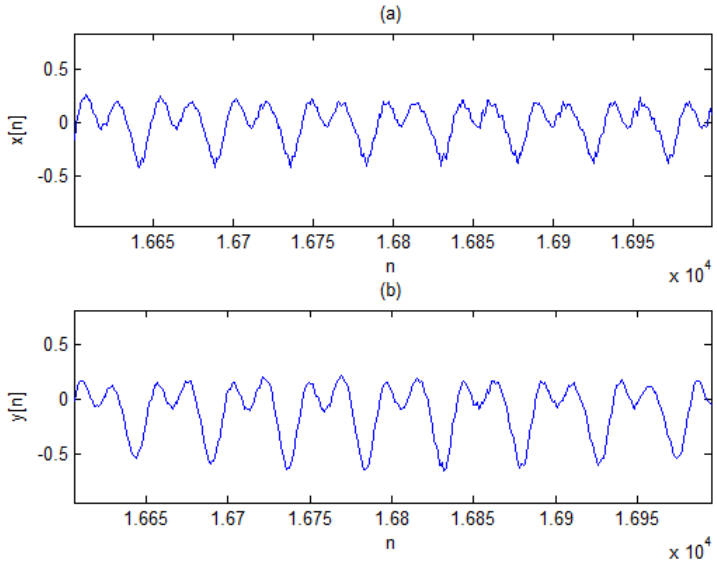

Figura 5.8: *(a) Señal original de voz femenina. Acercamiento a la forma de onda. (b) Versión sintética de (a) a* $10.9kbit/s$.

La figuras 5.8 y 5.9 presentan un acercamiento a las formas de onda de la señal original y la recuperada, en el fragmento mostrado de 5.8 las señales son muy parecidas mientras que en 5.9 al tener menor tasa de bits se pierde un poco la similitud.

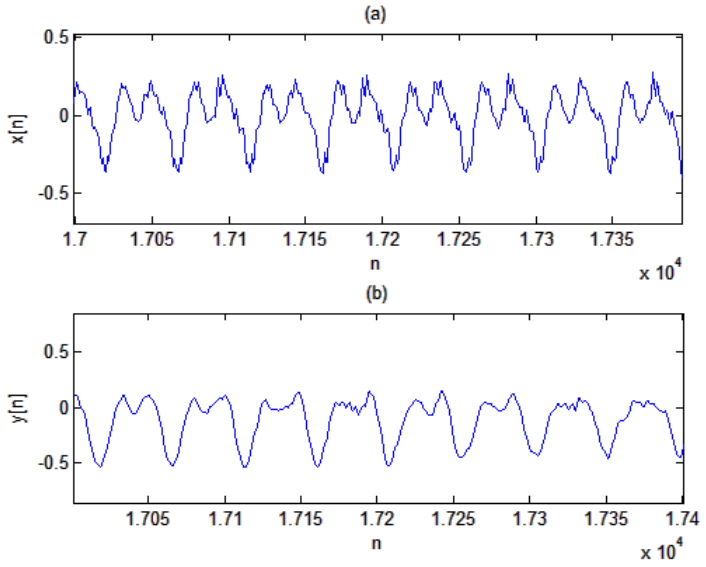

Figura 5.9: *(a) Señal original de voz femenina. Acercamiento a la forma de onda. (b) Versión sintética de (a) a* $7.8kbit/s$.

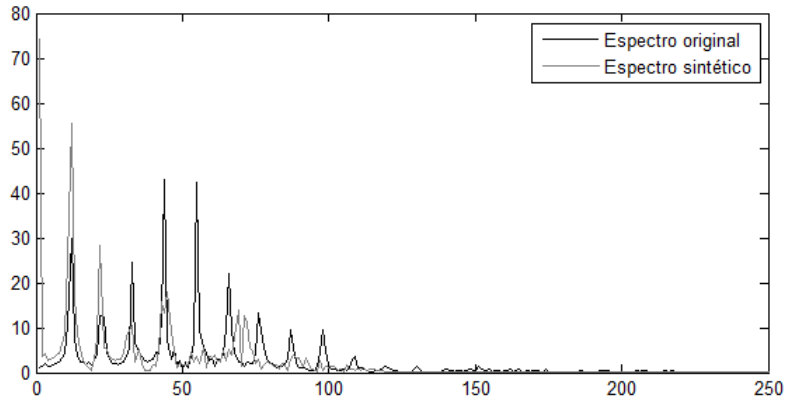

Figura 5.10: *Comparación de espectros del segmento de señal de la figura 5.7.*

Este codificador basa su funcionamiento en aproximar el espectro de la señal original para generar la señal sintética por lo que es más siginificativa la comparación de espectros de las señales.

Capítulo 6

Cancelación Adaptativa de Ruido

Existen diversas aplicaciones en el procesamiento digital de señales en las que los parámetros estadísticos no se pueden identificar a priori. En tales aplicaciones, como la cancelación de ruido, se emplean filtros con coeficientes ajustables denominados filtros adaptativos. Estos filtros adaptan sus coeficientes a los parámetros estadísticos de la señal mediante algoritmos.

Existen dos algoritmos básicos que calculan los coeficientes de un filtro adaptativo, el algoritmo de mínimos cuadrados (LMS, Least Mean Square) el cual pertenece a la familia de los algoritmos de gradiente estocástico, es decir, el filtro se adapta en base al error en el instante actual, y el algoritmo recursivo de mínimos cuadrados (RLS, Recursive Least Square) [20].

Los filtros adaptativos se han utilizado en sistemas en los que las características estadísticas de las señales que deben ser filtradas se desconocen, o en las que son variantes en el tiempo (señales no estacionarias). Tanto los filtros FIR como los IIR son considerados para el filtrado adaptivo, pero el filtro FIR es mucho más práctico y ampliamente utilizado. La razón es que a este filtro sólo hay que ajustarle los ceros y no presentan el problema de estabilidad asociado a los IIR, que tienen ceros y polos ajustables. Sin embargo, los filtros FIR adaptativos no siempre son estables, pues depende del algoritmo utilizado para ajustar sus coeficientes [11].

6.1. Filtros Adaptativos con el Algoritmo LMS

El principio de un algoritmo de filtro adaptativo se basa en la minimización de la diferencia entre la salida del filtro y una señal de referencia. Los coeficientes del filtro se actualizan de acuerdo al algoritmo LMS para que el error sea minimizado. Dicho algoritmo es frecuentemente utilizado gracias a su poca complejidad computacional.

El objetivo de un filtro adaptativo, dada una señal de entrada $x[n]$ y una señal deseada $d[n]$, es encontrar los coeficientes del filtro tal que su salida $y[n]$ se parezca lo más posible a la señal $d[n]$. Para conseguirlo se calcula el error $e[n] = d[n] - y[n]$ para ser minimizado con el algoritmo elegido.

Para formular el problema se dispone de un sistema desconocido con forma de filtro FIR

$$y[n] = \sum_{k=0}^{M} h[k]x[n-k] \tag{6.1.1}$$

donde $h[k]$ son los M coeficientes del filtro, $x[n]$ es la señal de excitación [26].

La secuencia de error se genera con

$$e[n] = d[n] - \sum_{k=0}^{M} h[k]x[n-k] \tag{6.1.2}$$

El algoritmo LMS encuentra los coeficientes del filtro que permiten obtener el valor esperado mínimo del cuadrado de la señal de error, y se deriva de la relación que guarda con el algoritmo de máximo descenso. Se sustituye el promedio estadístico $E[e(n)x(n)]$ por su valor instantáneo $e(n)x(n)$. Cada coeficiente del filtro muestra un comportamiento ruidoso con lo que la trayectoria de los coeficientes pasa de ser determinista a ser aleatoria. No existe garantía de que se alcance el filtro óptimo, sino que después de cierto número de iteraciones se moverá ruidosamente en torno al punto del error mínimo.

Partiendo del método de máximo descenso que establece que [2]

$$W[n+1] = W[n] - \mu \triangle J(W[n]) \tag{6.1.3}$$

$$\triangle J(W[n]) = -2r_{dx} + 2R_x W[n] \tag{6.1.4}$$

donde $\triangle J(W[n])$ es el vector gradiente definido como la derivada del error cuadrático medio con respecto al vector de coeficientes $W[n]$, R_x es la matriz de correlación y r_{dx} la correlación cruzada entre d y x.

Para el algoritmo LMS se cambian los estimadores R_x y r_{dx} por los estimados instantáneos

$$R_u \cong x[n]x^T[n] \tag{6.1.5}$$

y

$$r_{dx} \cong d[n]x[n] \tag{6.1.6}$$

Al sustituir (6.1.5) y (6.1.6) en (6.1.3) y (6.1.4) se obtiene

$$W[n+1] = W[n] + 2\mu x[n]\left(d[n] - x^T[n]W[n]\right) \tag{6.1.7}$$

$$W[n+1] = W[n] + 2\mu x[n]\left(d[n] - W^T[n]x[n]\right) \tag{6.1.8}$$

$$W[n+1] = W[n] + 2\mu e[n]x[n]. \tag{6.1.9}$$

La salida del filtro es

$$y[n] = w^T[n]x[n] \tag{6.1.10}$$

y la señal de error

$$e[n] = d[n] - y[n] \tag{6.1.11}$$

Los coeficientes del filtro en el tiempo n son

$$w[n] = [w_0[n]\ w_1[n]\ \ldots w_{M-1}[n]]^T \tag{6.1.12}$$

y la entrada del filtro [2]

$$x[n] = [x[n]\ x[n-1]\ x[n-2]\ \ldots x[n-M+1]]^T \tag{6.1.13}$$

Las ecuaciones (6.1.9), (6.1.10) y (6.1.11) constituyen el algoritmo adaptativo LMS. El parámetro μ controla la velocidad de estabilidad y convergencia del algoritmo [26].

6.2. Sistema de Cancelación de Ruido del Entorno en Señales de Voz

El sistema de cancelación de ruido se desarrolló para asistir al sistema de codificación de voz entregando señales de voz libres de ruido, o con el ruido atenuado cuando dicho sistema fuera utilizado en ambientes ruidosos. Para que el sistema funcione, de acuerdo a lo establecido en los textos, necesita de un algoritmo adaptable y elementos de hardware tales como un dispositivo para capturar audio con dos canales (convertidor analógico digital) y dos micrófonos de características similares. Un micrófono se utiliza para captar la voz junto con el ruido del entorno y el otro sólo debe captar el ruido del entorno para tomarlo como referencia de lo que debe cancelar. Es preferible que este último no capte, o lo haga lo menos posible, la voz dirigida al primer micrófono.

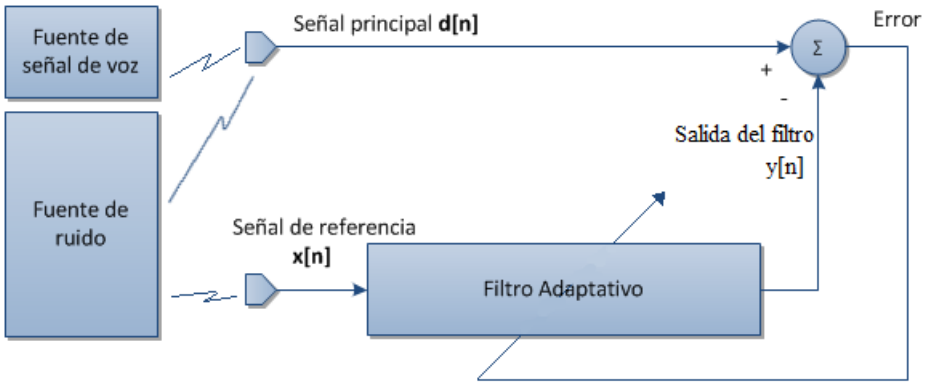

Figura 6.1: *Sistema adaptativo de cancelación de ruido*

6.2.1. Preprocesamiento

Las señales de entrada son divididas en tramas con ventana rectangular y son filtradas con

$$H(z) = \frac{1 - z^{-1}}{1 - 0.999z^{-1}} \tag{6.2.1}$$

el cual es un filtro que elimina el componente de corriente directa que puede ser introducido por el dispositivo de captura de audio.

6.3. Aplicación del Filtro Adaptativo en la Cancelación de Ruido en Señales de Voz

Para aplicar el filtro adaptativo a una aplicación real en la que no se cuenta con la señal deseada, la configuración del modelo se muestra en la figura 6.1. El primer micrófono debe captar la señal principal compuesta por la señal de voz s y la interferencia n_0, quedando como $d = s + n_0$ ocupando el lugar de la señal deseada. El segundo micrófono debe captar la interferencia n_1 la cual no debe estar correlacionada con la señal s pero si con n_0 de alguna manera. Este micrófono provee la señal de referencia al sistema. El ruido n_1 es filtrado para producir una salida y que debe ser una réplica de n_0. Esta salida es sustraída de la señal principal $s + n_0$ para producir la salida del sistema $s + n_0 - y$ [12] .

Para la captura de señales e implementación del algoritmo se deben considerar algunos factores [7] :

- El espectro del ruido de salida depende del espectro del ruido de la entrada, lo cual es fácil de aceptar.

- Si la SNR de la señal de referencia es alta, el ruido a la salida será bajo; es decir, mientras más pequeños sean los componentes de la señal s en la entrada de referencia, mejor será la cancelación del ruido a la salida del sistema.
- Si la SNR de la entrada principal es alta, el entrenamiento del filtro será más eficiente para cancelar el ruido n_0 en lugar de la señal s y en consecuencia el ruido a la salida será bajo.

6.3.1. Procedimiento

La división en tramas se hace con traslape para correlacionar los extremos finales de las tramas con los iniciales y reducir el ruido de los sonidos molestos creados por una concatenación de datos no correlacionados. Se selecciona una longitud de trama L y una longitud de traslape OL de tal manera que cuando se toma la trama actual, se le añade a su inicio, la última parte de longitud OL de la trama anterior y se añade a su final, la primera parte de longitud OL de lo que será la siguiente trama, con lo que se crea un vector de longitud $L + 2OL$, el cual es enviado a procesar. Para reconstruir la señal mediante la concatenación de las tramas filtradas, primero se crea una ventana de Hanning $ha[n]$ de longitud $4OL$. Se toma la primera mitad de la ventana ($ha[1 : 2OL]$) y se multiplica por el *Traslape a* de la trama actual de salida y por la región inicial de longitud OL de la misma trama, formando un vector de longitud $2OL$. Luego se toma la segunda mitad de la ventana $ha[n]$ ($ha[2OL + 1 : 4OL]$) y se multiplica por la última región de longitud OL de la trama de salida anterior y por el *Traslape b* de la misma trama, formando un segundo vector de longitud $2OL$. Finalmente se suman estos dos vectores para concatenar la última trama de salida con la anterior.

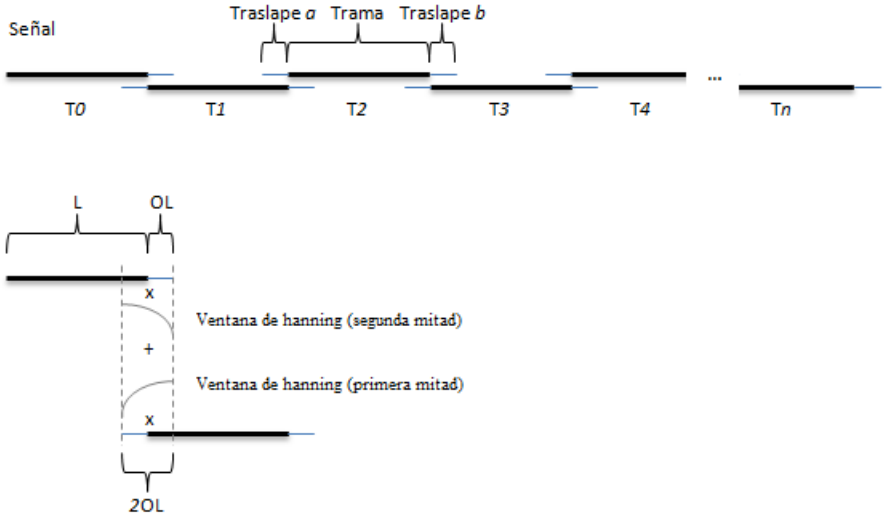

Figura 6.2: *Forma de dividir la señal en tramas y concatenación.*

La intervención de la ventana de Hanning evita que, al sumar las partes de traslape, se eleve la magnitud de la señal en esa región y existan saltos semejantes a un escalón, lo cual introduce ruido. Al sumar la primera mitad de $ha[n]$ con su segunda mitad se mantiene la unidad, lo que le da uniformidad a la señal reconstruida y suaviza las partes donde se concatenan las tramas.

Las señales $x[n]$ y $d[n]$ son divididas en tramas, como se explicó anteriormente para entrar al algoritmo adaptativo cuya salida y será de la misma longitud que x y d.

Aún cuando la entrada principal está compuesta de dos señales, una de información s y otra de interferencia n_0, ésta se coloca en el lugar de la señal deseada gracias al grado de correlación que guardan la interferencia de la entrada principal y la entrada de referencia. Esto permite a la señal de referencia parecerse a n_0 antes de que empiece a transformarse en d completamente, puesto que es la señal deseada.

6.3.2. Resultados con el Algoritmo Teórico Aplicado

Al aplicar el filtro adaptativo con el modelo mostrado en la figura 6.1 surgen algunos problemas y se observan algunas características no contempladas antes. El más notorio problema es el uso de la resta aritmética para sustraer el ruido y de la entrada principal $s + n_0$, puesto que es muy necesario que y se aproxime al 100 % a n_0 para que al restar se elimine el ruido en su totalidad; sin embargo, en la mayoría de los casos, el ruido a cancelar tiene un contenido frecuencial muy variado pero abundante en altas frecuencias, lo que implica la presencia de información detallada que, al no ser perfectamente aproximada por el filtro, al aplicar la resta el resultado no favorece cancelación alguna sino que distorsiona la interferencia.

Aproximar detalles o altas frecuencias significa someter al algoritmo a un número mayor de iteraciones, dadas las características del filtro adaptativo, las bajas frecuencias se aproximan primero y rápidamente, las altas lo hacen lentamente y al final. Un factor, que influye en el tiempo de convergencia del algoritmo, es la calidad de las señales entregadas por los micrófonos. Es muy importante que ambos sean iguales para que las formas de onda de la interferencia sean lo más parecido y así estén lo más correlacionadas posible. El filtro adaptativo tiene la capacidad de converger aún cuando las señales presentan un defasamiento entre ellas, pero el tiempo requerido es mucho mayor, y al no incrementarlo, se pierde considerablemente el buen desempeño del sistema.

Es importante conocer las características del sistema de hardware y corregir defasamientos o inversiones en amplitud para mejorar el desempeño del sistema de cancelación. En un experimento realizado en el que fueron grabadas dos señales simultáneamente, la señal principal y la referencia, se encontró que la señal principal estaba adelantada treinta muestras respecto a la otra, además estaba invertida respecto al eje del tiempo. Esto provocó un mal desempeño en la cancelación por lo que se procedió a corregir la posición de las señales para favorecer la convergencia del sistema.

Otro problema surge cuando, al ser la entrada principal la señal deseada, la salida y comienza a aproximarse a la voz s, provocando además de la cancelación del ruido, el comienzo de la cancelación de la voz dejándola distorsionada. No obstante, se encontró que, aunque parece que y se aproxima a la voz y su envolvente es muy parecida a la de la entrada principal d con voz y ruido, las regiones de y que corresponden a las regiones en d que tienen voz, contienen información correspondiente al ruido más que a la voz. Lo que ocurre es que en dichas regiones y adquiere las amplitudes de d y una forma de onda similar pero la información corresponde al ruido. Lo anterior se soluciona evitando que y crezca a niveles de energía superiores a los de la señal de referencia en la misma región.

6.3.3. Resultados

La SNR de la señal principal es $SNR = 21dB$ y la SNR de la salida del sistema es $SNR = 35dB$. El incremento puede considerarse bueno y significativo. Al escuchar la señal de salida se aprecia una disminución importante del ruido comparado con el de la señal principal.

En la figura 6.3 se muestra la gráfica de las señales de referencia y principal utilizadas para probar el algoritmo. Las señales tienen una duración de 3 segundos y $fs = 10kHz$.

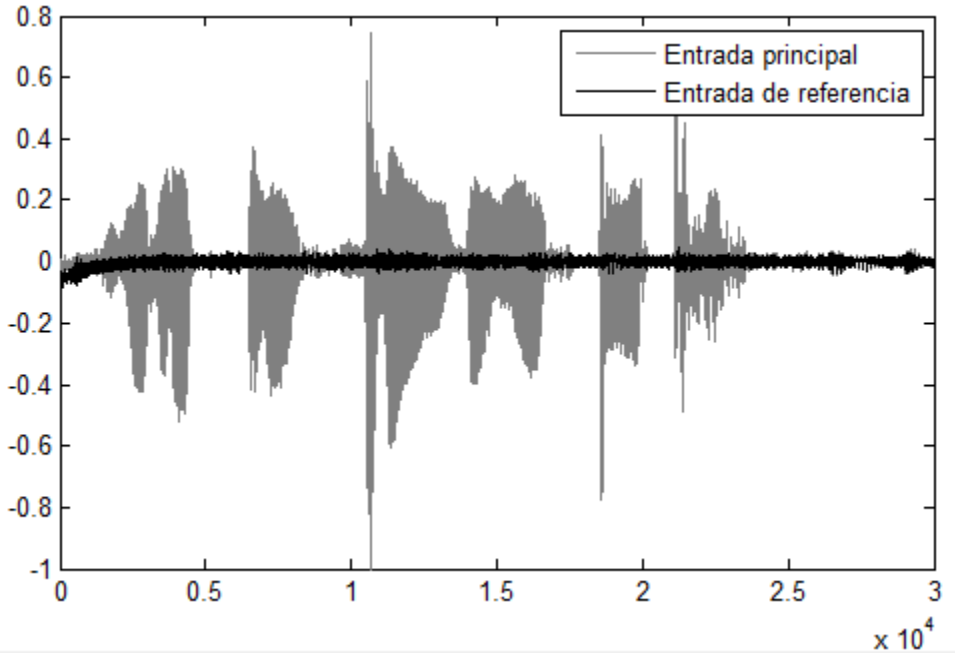

Figura 6.3: *Entrada principal y de referencia al filtro adaptativo*

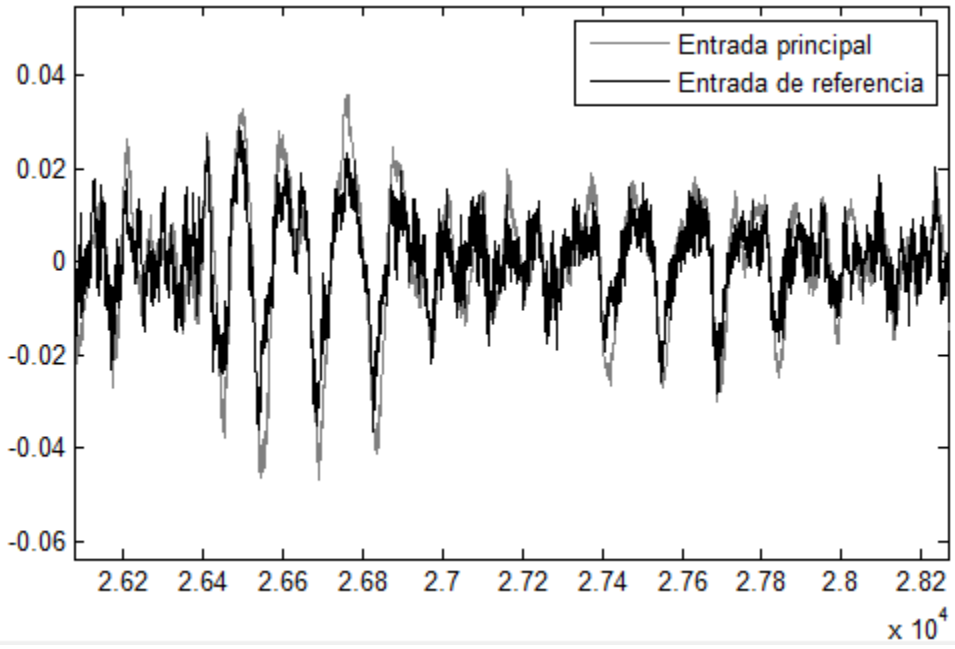

Figura 6.4: *Vista aumentada de las señales de entrada al filtro adaptativo.*

En el acercamiento a las señales mostrado en la figura 6.4 se aprecia el grado de correlación que guardan ambas señales, sus formas de onda son parecidas entre sí; sin embargo, no es suficiente ese grado de similitud

para realizar la sustracción aritmética.

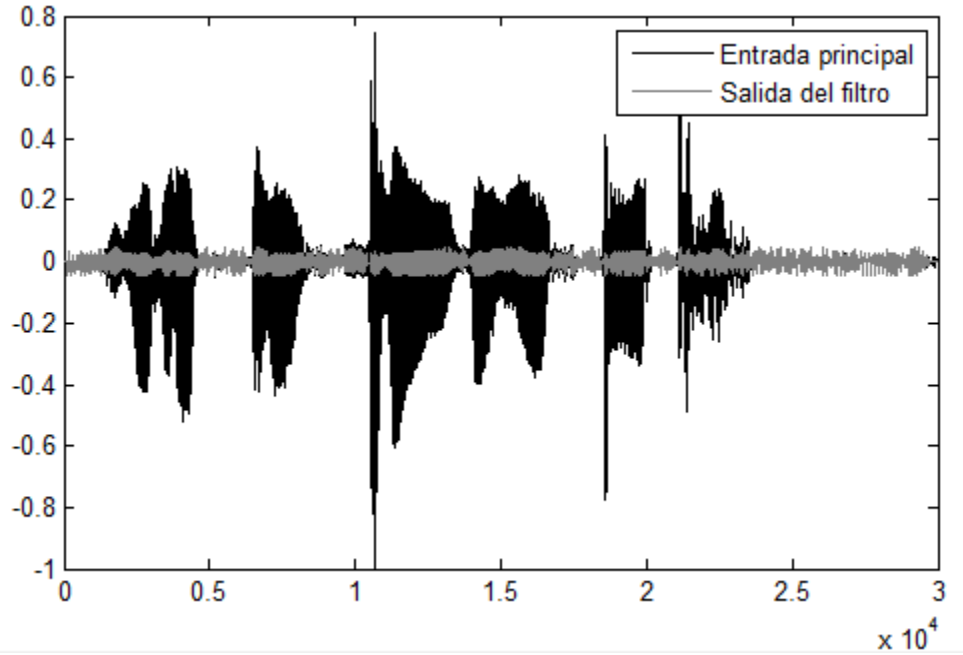

Figura 6.5: *Entrada principal y salida del filtro adaptativo.*

En la figura 6.5 se observa la salida y del filtro adaptativo la cual debe ser igual al componente de ruido de la señal principal. En las figuras 6.6 y 6.7 se muestra un acercamiento que permite apreciar la similitud entre las señales de ruido n_0 y y.

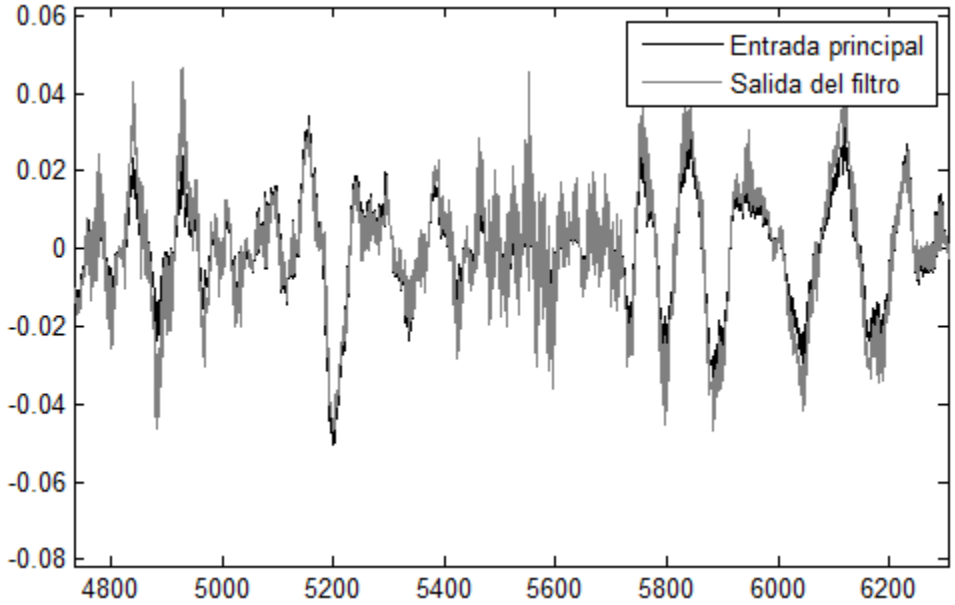

Figura 6.6: *Vista aumentada de la entrada principal y salida del filtro.*

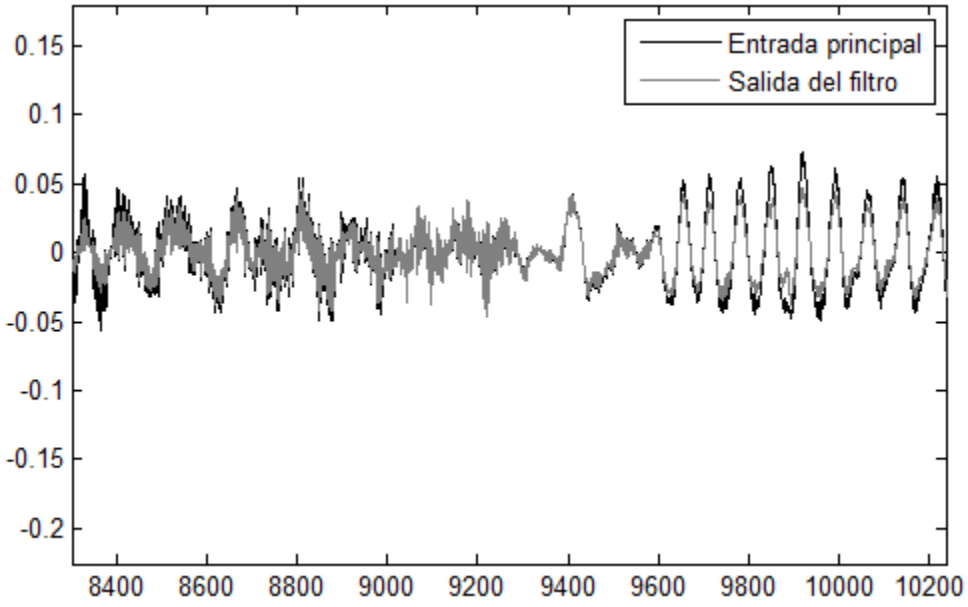

Figura 6.7: *Vista aumentada de la entrada principal y salida del filtro. Otra región de las señales*

Ahora se muestra, en figura 6.8 la salida del sistema que corresponde a la resta $s + n_0 - y$ y debe resultar en s, la señal sin ruido.

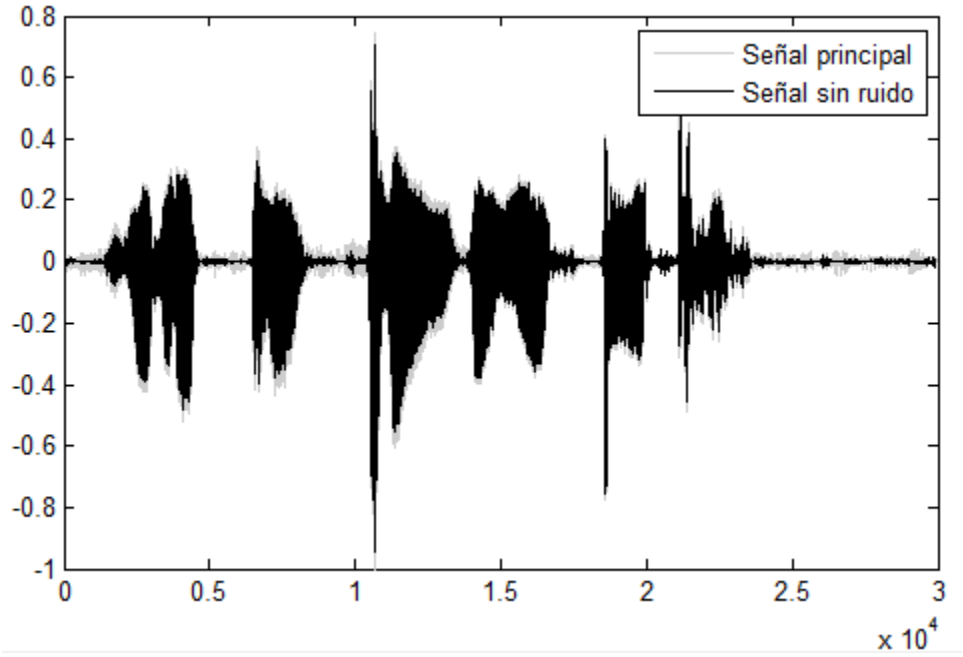

Figura 6.8: *Entrada principal y salida del sistema.* $\mu = 0.1$.

La gráfica de color negro muestra la señal sin ruido y atrás de ésta la señal principal sobresaliendo en amplitud por los componentes de ruido.

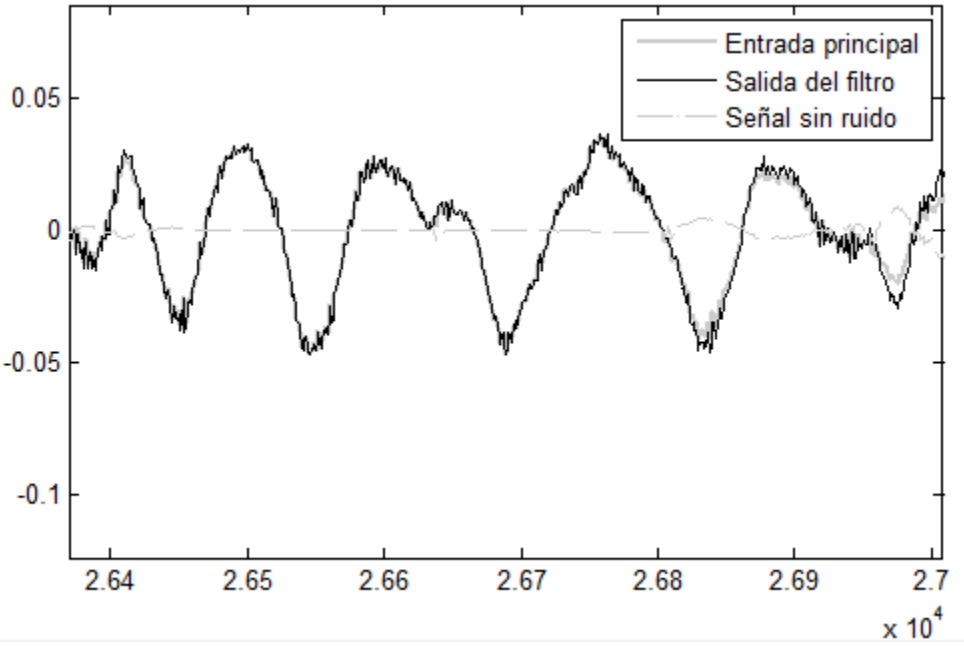

Figura 6.9: *Muestra de la similitud de* y *con* n_0 *y la resta entre éstas.*

El resultado de la cancelación de ruido es bueno con un incremento de $14dB$ en la SNR. Por otro lado, la implementación de este algoritmo en tiempo real bajo las condiciones de hardware en este experimento y la sencillez del modelo teórico no es factible dado que por cada trama se requiere de 1000 a 3000 iteraciones para tener un buen desempeño. Esto genera un tiempo de procesamiento superior al de la duración de una trama.

6.4. Solución Alterna

Cancelar ruido cuyas características se desconocen implica el uso de un sistema adaptativo, donde es necesario tener una señal de referencia de lo que se desea cancelar. No obstante, existe la posibilidad de eliminar el ruido de manera muy eficiente, con poco procesamiento y sin la entrada de referencia. El sistema alterno contiene los mismos elementos del filtro adaptativo definido anteriormente, pero ahora sólo utiliza una entrada. La entradas principal y de referencia son la misma, reciben la misma señal, la cual está compuesta por las señales de voz y de ruido como se muestra en la figura 6.10. Lo anterior es posible dadas las características de los coeficientes w del filtro, del ruido de la señal y del comportamiento del filtro al variar μ. La SNR de la señal principal debe ser alta para mejorar el desempeño del sistema.

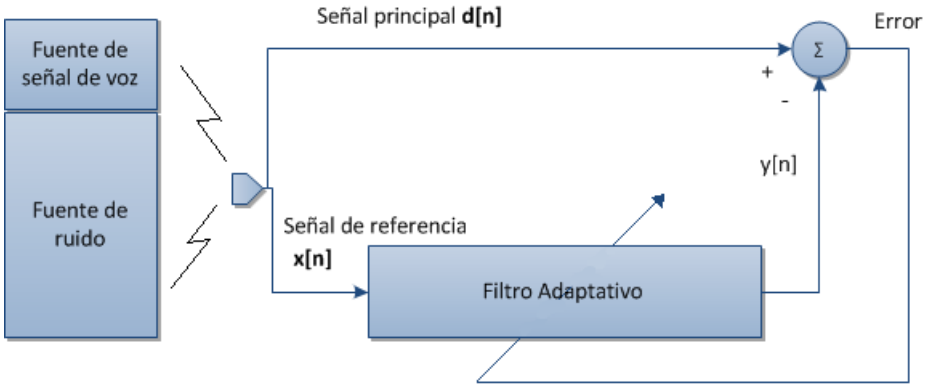

Figura 6.10: *Filtro adaptativo que utiliza la misma señal como principal y referencia.*

Conforme crece μ, el algoritmo converge con mayor rapidez. Los elementos de mayor energía son aproximados primero y si la SNR de la entrada principal es alta, se reconstruirá primero la señal de voz al tener mayor energía. Otro aspecto importante es que las bajas frecuencias suelen tener mayor energía que las altas frecuencias, lo que provoca que la señal de voz sin ruido se perciba con un efecto de haber sido filtrada con un filtro paso bajas. Ciertamente, la respuesta en frecuencia de los pesos w que modelan el filtro adaptativo tienen tendencia paso bajas. En las figuras 6.11 y 6.12 se observa su comportamiento al variar μ.

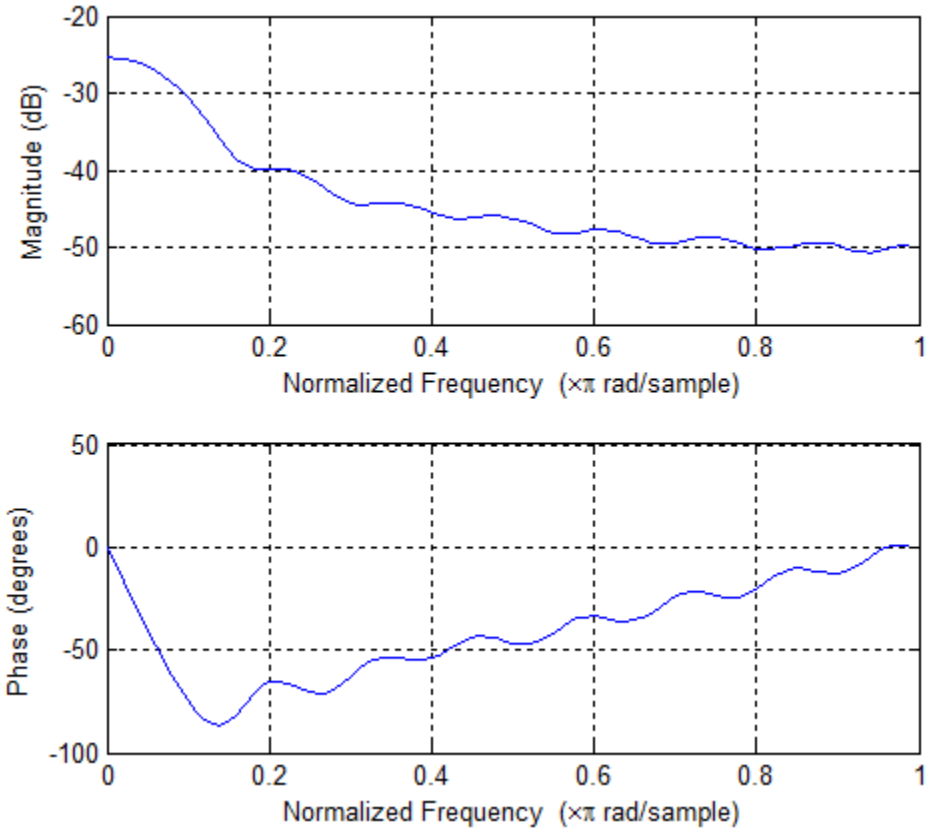

Figura 6.11: *Respuesta en frecuencia de los pesos w del filtro adaptativo cuando $\mu = 0.001$.*

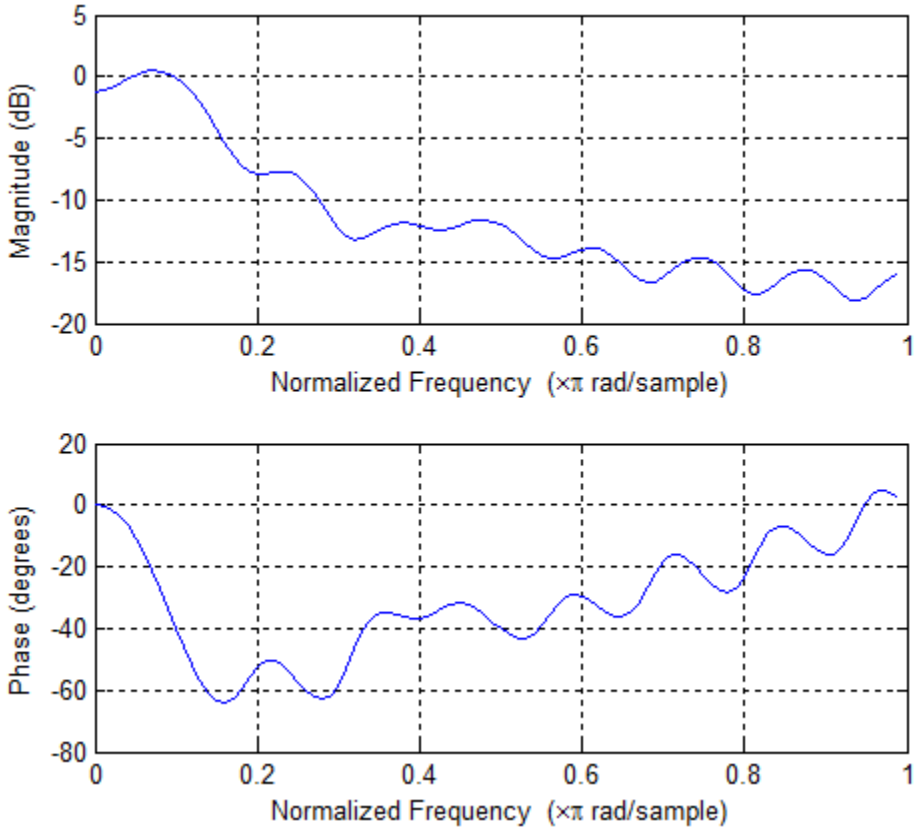

Figura 6.12: *Respuesta en frecuencia de los pesos w del filtro adaptativo cuando* $\mu = 0.2$.

Al incrementar μ, aumenta la magnitud de la respuesta en frecuencia del filtro, lo que limita la aparición de la parte de ruido en la salida cuando su nivel de energía es lo suficientemente bajo para no ser aproximado pero si lo suficiente para ser escuchado y tener la necesidad de cancelar el ruido.

El desempeño de este modelo es muy bueno y lo favorece el uso de un micrófono unidireccional que al tener la fuente de la información principal relativamente cerca y la fuente de interferencia relativamente lejos, la energía de la primera será mucho mayor que la segunda. La energía de la fuente de interferencia será baja por dos razones: la relativa lejanía con respecto a la fuente de voz o de información principal, y la característica unidireccional del micrófono que no le permite captar de manera favorable sonidos provenientes de fuentes alejadas y de diferentes direcciones.

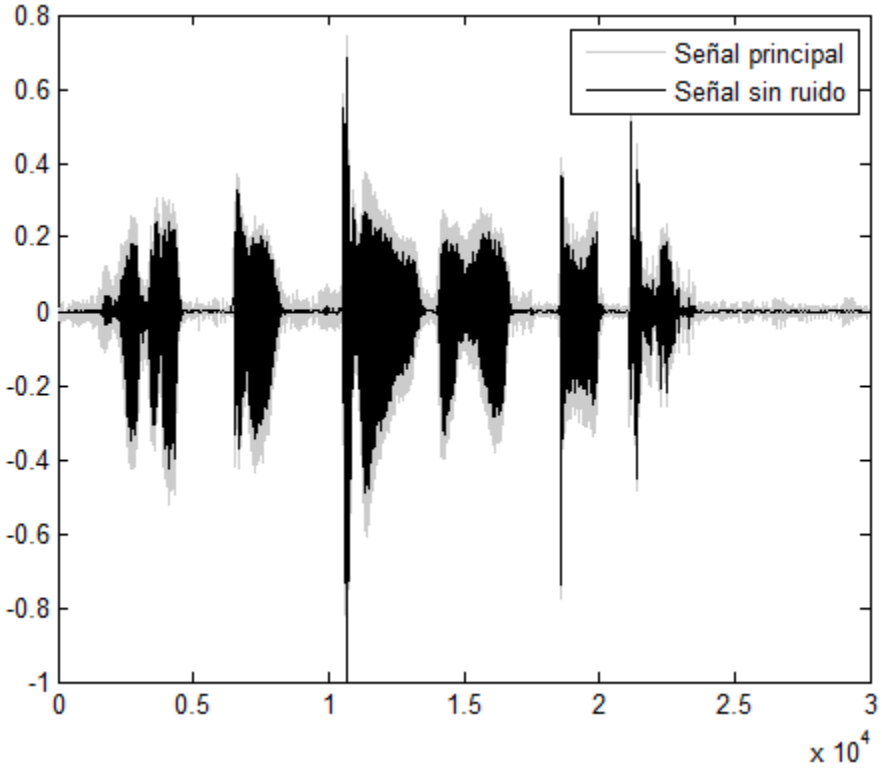

Figura 6.13: *Salida sin ruido y señal principal.* $\mu = 0.08$.

En la figura 6.13 se muestra la salida sin ruido con una $SNR = 53dB$, esto es, $32dB$ mayor que la señal principal. El valor de μ seleccionado es $\mu = 0.08$. El ruido no se percibe en absoluto en la señal de salida y, sólo presenta el inconveniente de haber perdido un poco de información de detalles, la cual no es imprescindible para la inteligibilidad del mensaje.

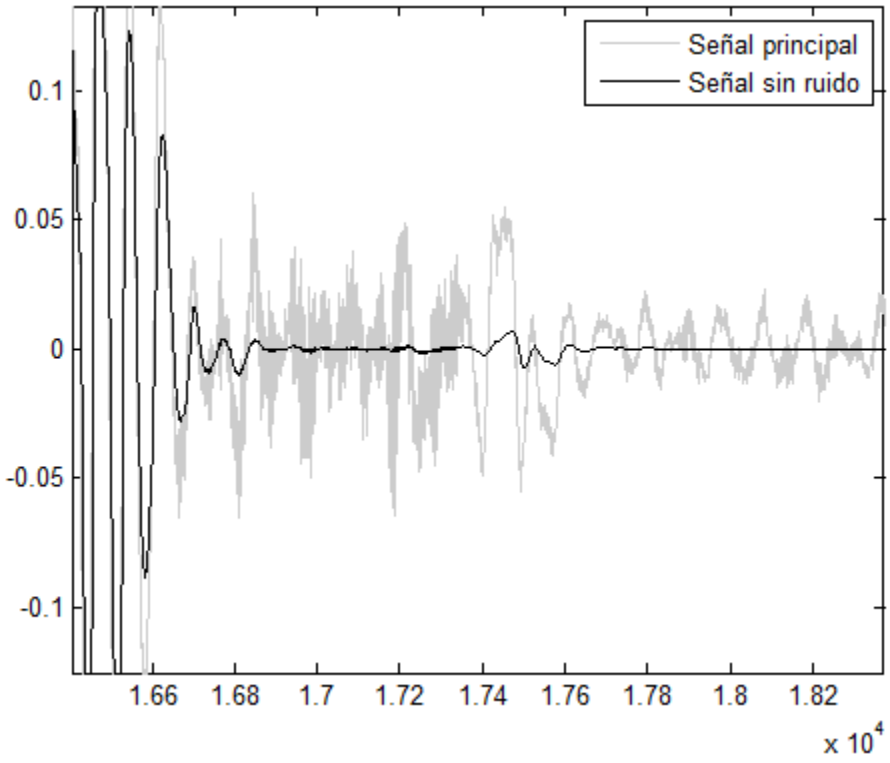

Figura 6.14: *Acercamiento de la señal sin ruido y la señal principal.*

En la figura 6.14 que muestra el acercamiento se observa la atenuación considerable del ruido. Si se utiliza un valor de μ más bajo se comprueba la lenta aproximación de la señal en el filtro; ver figura 6.15.

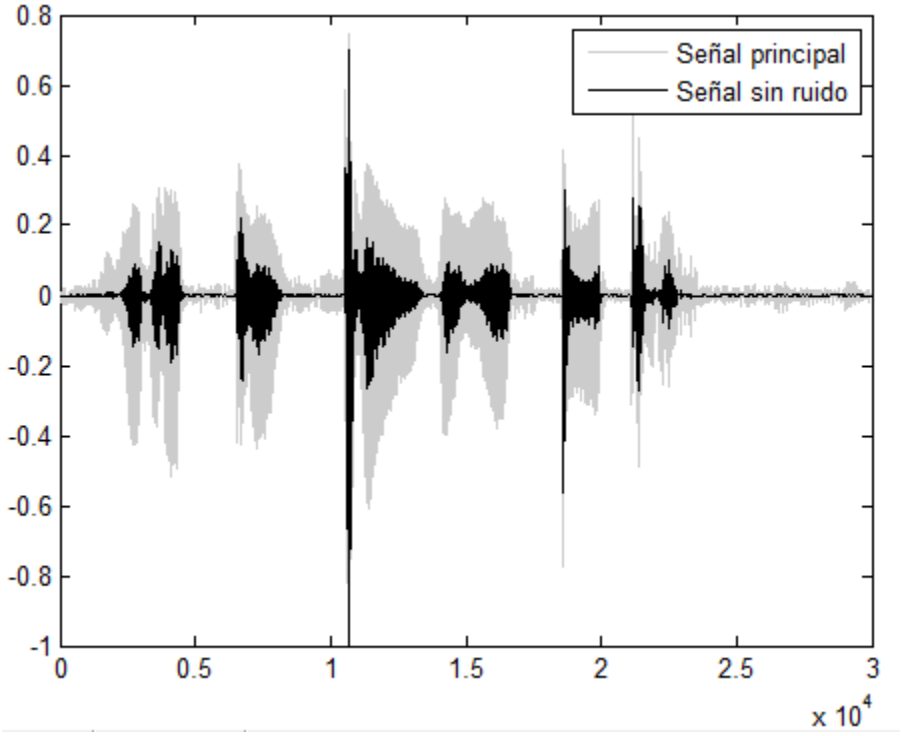

Figura 6.15: *Salida sin ruido y señal principal.* $\mu = 0.008$.

Este modelo no requiere de un gran número de iteraciones y puede ser implementado en tiempo real. La cancelación de ruido es notable y puede ser utilizado en diversas aplicaciones.

Capítulo 7

Conclusiones

Las diferencias entre las escalas mel y bark son casi imperceptibles en las condiciones de los ejemplos mostrados en el Capítulo 5. Cuanto más se incrementa la tasa de bits, las diferencias son completamente imperceptibles. Para una muy baja tasa de bits existen algunas diferencias audibles entre las escalas, pero no son estadísticamente significativas.

La calidad de la señal sintética es realmente buena, el desempeño de codificador es muy bueno puesto que la presencia de ruido no le afecta. El decodificador es capaz de reconstruir cualquier tipo de ruido aunque no sea con la misma calidad dada a las señales de voz.

La distorsión espectral de los parámetros después de un proceso de cuantización es muy baja. Cuando se hace la comparación entre una señal sintética generada con parámetros cuantizados y una señal sintética generada con parámetros no cuantizados casi no hay diferencias perceptibles.

El sistema propuesto se desempeña muy bien y es capaz de sintetizar música y ruido de la calle (automóviles, aviones, gente, etc.). No importa qué interfiere con la voz de alguien porque ésta es siempre inteligible después de haber sido comprimida y recuperada en este sistema de codificación.

El sistema está compuesto de diferentes técnicas de procesamiento de voz cuyo desempeño se ha probado en tareas específicas dentro del codificador. La compresión con wavelets tiene mejor desempeño con señales de análisis y como sólo una parte del sistema, que con señales de voz y como un único sistema de compresión.

A pesar de la sencillez del sistema en cuanto a etapas de procesamiento y en comparación con un sistema CELP, el desempeño es muy bueno por los resultados y por pensar en que se puede complementar con elementos utilizados en un codificador LPC para mejorar la calidad o disminuir la tasa de bits.

Se tiene ahora una base de codificador que involucra el análisis cepstral, wavelet y el uso de la señal de análisis que puede complementarse con diversos elementos que existen en los codificadores LPC y crear un codificador para una aplicación específica.

El sistema de cancelación adaptativa de ruido mostró muy buenos resultados pese a las modificaciones realizadas a los modelos más comunes y a la utilización de un solo micrófono. La relación señal a ruido de la señal de salida se incrementa considerablemente, con lo que se cuenta con un sistema de baja complejidad computacional (dado que sólo procesa una señal) que ofrece la reducción del ruido del entorno.

Apéndice A

Demostración del Origen de las Ecuaciones (4.2.2) y (4.2.3)

Sean

$$z = \frac{w+\alpha}{1+\alpha w} \text{ y } \alpha = \frac{w_3 - z_3}{z_3 w_3 - 1}$$

que se derivan de (4.2.1)

$$\frac{z-z_1}{z-z_3}\frac{z_2-z_3}{z_2-z_1} = \frac{w-w_1}{w-w_3}\frac{w_2-w_3}{w_2-w_1},$$

al sustiuir $z_1 = w_1 = 1$ y $z_2 = w_2 = -1$ que corresponden al mapeo de dc a dc y la frecuencia máxima a la frecuencia máxima.

Aplicando las sustituciones de z_1, w_1, z_2 y w_2

$$\frac{z-1}{-1-1}\frac{1-z_3}{z-z_3} = \frac{w-1}{-1-1}\frac{w_2-w_3}{w_2-w_1} \tag{A.0.1}$$

$$\frac{w-1}{-1-z_3}\cdot\frac{-1-w_3}{w-w_3} = \frac{z-1}{z-3} \tag{A.0.2}$$

$$\frac{(z-z_3)(w-1)(-1-w_3)}{(-1-z_3)(w-w_3)} = z-1 \tag{A.0.3}$$

$$z = \frac{(z-z_3)(w-1)(-1-w_3)}{(-1-z_3)(w-w_3)} + 1 \tag{A.0.4}$$

Sea $A = -1 - w_3$ y $B = -1 - z_3$ entonces

$$z = \frac{(z-z_3)(w-1)\cdot A}{B\cdot(w-w_3)} + 1 \tag{A.0.5}$$

$$z = \frac{Awz - Az - Awz_3 + Az_3}{B\cdot(w-w_3)} + \frac{B\cdot(w-w_3)}{B\cdot(w-w_3)} \tag{A.0.6}$$

$$Bzw - Bzw_3 = Azw - Az - Awz_3 + Az_3 + Bw - Bw_3 \tag{A.0.7}$$

$$z(Bw - Bw_3 - Aw + A) = -Awz_3 + Az_3 + Bw - Bw_3 \tag{A.0.8}$$

$$z = \frac{w(B - Az_3) + Az_3 - Bw_3}{w(B - A) - Bw_3 + A} \tag{A.0.9}$$

$$z = \frac{\frac{w(B-Az_3)}{B-Az_3} + \frac{Az_3-Bw_3}{B-Az_3}}{\frac{w(B-A)}{B-Az_3} + \frac{Bw_3+A}{-B-Az_3}} \tag{A.0.10}$$

$$z = \frac{w + \frac{Az_3-Bw_3}{B-Az_3}}{\frac{w(B-A)}{B-Az_3} + \frac{Bw_3+A}{-B-Az_3}} \tag{A.0.11}$$

Si

$$B - A = w_3 - z_3 \tag{A.0.12}$$

$$Az_3 - Bw_3 = w_3 - z_3 \tag{A.0.13}$$

$$B - Az_3 = w_3z_3 - 1 \tag{A.0.14}$$

$$A - Bw_3 = w_3z_3 - 1, \tag{A.0.15}$$

entonces

$$\frac{B - A}{B - Az_3} = \frac{Az_3 - Bw_3}{B - Az_3} = \frac{w_3 - z_3}{w_3z_3 - 1} = \alpha \tag{A.0.16}$$

$\therefore z = \frac{w+\alpha}{1+\alpha w}$

Apéndice B

Nota Media de Opinión. Prueba MOS

En sistemas de comunicaciones de voz, la calidad refiere a si la experiencia es buena o mala. Existe un método para evaluar la calidad del audio que utiliza una escala numérica en lugar de dar descripciones del tipo "muy bien", o "muy mal". La Nota media de opinión [22] (MOS, Mean Opinion Score) da una indicación numérica de la calidad percibida de la señal de audio después de haber sido comprimida.

La escala numérica de evaluación es la siguiente:

Tabla B.1: *Tabla de calificaciones MOS.*

MOS	Calidad	Deterioro
5	Excelente	Imperceptible
4	Buena	Perceptible pero no molesto
3	Regular	Un poco molesto
2	Pobre	Molesto
1	Mala	Muy molesto

La calificación 5 equivale a una conversación cara a cara o recepción de radio; con 4 se perciben imperfecciones pero el sonido es claro, es el rango de los celulares; en 3 se perciben imperfecciones poco molestas; en 2 existe dificultad para comunicarse y en 1 es imposible comunicarse [28].

Apéndice C

Aproximación de Padé

La aproximación de Padé es una técnica desarrollada por Henri Padé, la cual se define como la mejor aproximación de una función por una función racional de cierto orden, donde las series de potencias de los aproximantes coinciden con las series de potencias de la función que se está aproximando. La aproximación Padé ofrece una mejor aproximación de la función que su serie de Taylor truncada, además puede funcionar donde las series de Taylor no convergen [10].

Sea f una función, $m \geq 0$ y $n \geq 0$ dos enteros, la aproximación de Padé de orden $[m/n]$ o $N = m + n$ es la función racional

$$R(x) = \frac{P_m(x)}{Q_n(x)} = \frac{\sum_{j=0}^{m} a_j x^j}{1 + \sum_{k=1}^{n} b_k x^k} = \frac{a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m}{1 + b_1 x + b_2 x^2 + \cdots b_n x^n} \tag{C.0.1}$$

la cual coincide con $f(x)$ al más alto orden posible, lo que equivale a [10]
$f(0) = R(0)$
$f'(0) = R'(0)$
$f''(0) = R''(0)$
$\vdots$
$f^{(m+n)}(0) = R^{(m+n)}(0)$.

La aproximación de Padé es única para valores m y n dados, esto es, los coeficientes a_0, a_1, ..., a_m, $b_1, \ldots, b_n$ son determinados de manera única. Por razones de unicidad, se seleccionó $b_0 = 1$ para ser el término de orden cero en el denominador de $R(x)$. En consecuencia, hay $N + 1$ parámetros disponibles para aproximar f por medio de r.

La aproximación de Padé también se denota como

$$[m/n]f(x). \tag{C.0.2}$$

C.1. Cálculo de los Coeficientes

El método de aproximación de Padé selecciona los N+1 coeficientes de modo que $f^{(k)}(0) = f^{(k)}(0)$ para cada $k = 0, 1, \ldots, N$. La aproximación de Padé es una extensión de la aproximación polinomial de Taylor a funciones racionales. Además, cuando $n = N$ y $m = 0$, la aproximación de Padé es el N-ésimo polinomio de Maclaurin (Serie de Taylor centrada en cero).

Sea la diferencia

$$f(x) - R(x) = f(x) - \frac{P(x)}{Q(x)} = \frac{f(x)Q(x) - P(x)}{Q(x)} = \frac{f(x)\sum_{i=0}^{m} b_i x^i - \sum_{i=0}^{n} a_i x^i}{Q(x)}, \tag{C.1.1}$$

y suponiendo que $f(x)$ tiene la expansión de la serie de Maclaurin $f(x) = \sum_{i=0}^{\infty} c_i x^i$. Entonces

$$f(x) - r(x) = \frac{\sum_{i=0}^{\infty} c_i x^i \sum_{i=0}^{m} b_i x^i - \sum_{i=0}^{n} a_i x^i}{Q(x)}. \tag{C.1.2}$$

El objetivo es seleccionar las constantes b_1, b_2, $\ldots b_m$ y a_0, a_1, $\ldots$, a_n de tal forma que

$$f^{(k)}(0) - r^{(k)} = 0, \, para \, cada \, k = 0, \, 1, \ldots, \, N. \tag{C.1.3}$$

Posteriormente se selecciona b_1, b_2, $\ldots b_m$ y a_0, a_1, $\ldots$, a_n de manera que el numerador del lado derecho de la ecuación [8.14],

$$(a_0 + a_1 x + \cdots)\,(1 + q_1 x + \cdots + q_m x^m) - (p_0 + p_1 x + \cdots p_n x^n)\,, \tag{C.1.4}$$

no tenga términos de un grado menor o igual que N.

Con el fin de simplificar la notación se define $a_{n+1} = a_{n+2} = \cdots = a_N$ y $b_{m+1} = b_{m+2} = \cdots = b_N = 0$. Ahora se puede escribir el coeficiente de x^k en la expresión [8.15] como

$$\left(\sum_{i=0}^{k} c_i b_{k-1} \right) - a_k. \tag{C.1.5}$$

Entonces, la función racional de la aproximación de Padé proviene de la solución de las $N+1$ ecuaciones lineales:

$$\sum_{i=0}^{k} c_i b_{k-1} = a_k, \; k = 0, \, 1, \ldots, \, N \tag{C.1.6}$$

en las $N+1$ incógnitas b_1, b_2, $\ldots b_m$ y a_0, a_1, $\ldots$, a_n.

C.1.1. Ejemplo

La aproximación de Padé de e^{-x} se muestra a continuación. El desarrollo de la serie de Maclaurin para la exponencial es

$$\sum_{i=0}^{\infty} \frac{(-1)^i}{i!} x^i. \tag{C.1.7}$$

Para encontrar la aproximación de Padé a e^{-x} de quinto grado con $n = 3$ y $m = 2$ se requiere seleccionar a_0, a_1, a_2, a_3, b_1 y b_2 de manera que los coeficientes de x^k para $k = 0, \, 1, \ldots, \, 5$ sean cero en la expresión

$$\left(1 - x + \frac{x^2}{2} - \frac{x^3}{6} + \cdots\right)\left(1 + b_1 x + b_2 x^2\right) - \left(a_0 + a_1 x + a_2 x^2 + a_3 x^3\right). \tag{C.1.8}$$

Al expandir y agrupar los términos se obtiene

$x^5 : -\frac{1}{120} + \frac{1}{24} b_1 - \frac{1}{6} b_2 = 0;$

$x^4 : \frac{1}{24} - \frac{1}{6} b_1 + \frac{1}{2} b_2 = 0;$

$x^3 : -\frac{1}{6} + \frac{1}{2} b_1 - b_2 = a_3;$

$x^2 : \frac{1}{2} - b_1 + b_2 = a_2;$

$x^1 : -1 + b_1 = a_1;$

$x^0 : 1 = a_0$.

El sistema de ecuaciones se resuelve con algún programa de computadora. Las soluciones son $a_0 = 1$, $a_1 = -\frac{3}{5}$, $a_2 = \frac{3}{20}$, $a_3 = -\frac{1}{60}$, $b_1 = \frac{2}{5}$ y $b_2 = \frac{1}{20}$.

La aproximación de la función exponencial mediante Padé es común por lo que existen tablas elaboradas con los coeficientes de los polinomios P y Q de la función racional r para un orden $[m/n]$ dado.

Tabla C.1: *Coeficientes para la aproximación de la función exponencial*

m / n	0	1	2	3
0	$\frac{1}{1}$	$\frac{1}{1-x}$	$\frac{1}{1-x+\frac{1}{2}x^2}$	$\frac{1}{1-x+\frac{1}{2}x^2-\frac{1}{6}x^3}$
1	$\frac{1+x}{1}$	$\frac{1+\frac{1}{2}x}{1-\frac{1}{2}x}$	$\frac{1+\frac{1}{3}x}{1-\frac{2}{3}x+\frac{1}{6}x^2}$	$\frac{1+\frac{1}{4}x}{1-\frac{3}{4}x+\frac{1}{4}x^2-\frac{1}{24}x^3}$
2	$\frac{1+x+\frac{1}{2}x^2}{1}$	$\frac{1+\frac{2}{3}x+\frac{1}{6}x^2}{1-\frac{1}{3}x}$	$\frac{1+\frac{1}{2}x+\frac{1}{12}x^2}{1-\frac{1}{2}x+\frac{1}{12}x^2}$	$\frac{1+\frac{2}{5}x+\frac{1}{20}x^2}{1-\frac{3}{5}x+\frac{3}{20}x^2-\frac{1}{60}x^3}$
3	$\frac{1+x+\frac{1}{2}x^2+\frac{1}{6}x^3}{1}$	$\frac{1+\frac{3}{4}x+\frac{1}{4}x^2+\frac{1}{24}x^3}{1-\frac{1}{4}x}$	$\frac{1+\frac{3}{5}x+\frac{3}{20}x^2+\frac{1}{60}x^3}{1-\frac{2}{5}x+\frac{1}{20}x^2}$	$\frac{1+\frac{1}{2}x+\frac{1}{10}x^2+\frac{1}{120}x^3}{1-\frac{1}{2}x+\frac{1}{10}x^2-\frac{1}{120}x^3}$
4	$\frac{1+x+\frac{1}{2}x^2+\frac{1}{6}x^3+\frac{1}{24}x^4}{1}$	$\frac{1+\frac{4}{5}x+\frac{3}{10}x^2+\frac{1}{15}x^3+\frac{1}{120}x^4}{1-\frac{1}{5}x}$	$\frac{1+\frac{2}{3}x+\frac{1}{5}x^2+\frac{1}{30}x^3+\frac{1}{360}x^4}{1-\frac{1}{3}x+\frac{1}{30}x^2}$	$\frac{1+\frac{4}{7}x+\frac{1}{7}x^2+\frac{2}{105}x^3+\frac{1}{840}x^4}{1-\frac{3}{7}x+\frac{1}{14}x^2-\frac{1}{210}x^3}$

Apéndice D

Retardo de Grupo

A una señal le toma tiempo pasar a través de un sistema. El tiempo finito requerido a la señal para pasar por un filtro es llamado *retardo*.

D.1. Fase contra Frecuencia

Si el tiempo de tránsito de la señal por el sistema es el mismo para todas las frecuencias, se dice que la fase es lineal respecto a la frecuencia. Si el tiempo de tránsito es diferente en diferentes frecuencias, el resultado es una fase no lineal. Si la fase cambia no linealmente con la frecuencia, la señal de salida estará distorsionada.

La distorsión de retardo se expresa usualmente en unidades de tiempo: milisegundos (ms), microsegundos (μs), o nanosegundos (ns) relativa a la frecuencia de referencia.

El retardo de grupo se define como la derivada de la fase respecto a la frecuencia en radianes [15].

$$GD = \frac{d\varphi}{d\omega} \tag{D.1.1}$$

donde GD es la variación de retardo de grupo, φ es la fase en radianes y ω es la frecuencia en radianes por segundo.

Si la relación de la fase contra la frecuencia es no lineal, existe el retardo de grupo. En un sistema sin retardo de grupo, todas las frecuencias son transmitidas a través del sistema en la misma cantidad de tiempo, es decir; con igual tiempo de retardo. Si el retardo de grupo existe, algunas frecuencias viajan más rápido que otras.

Se puede hacer una analogía con una pista de atletismo sobre la cual corren atletas en sus respectivos carriles y en una carrera particular llegan al mismo tiempo. En un sistema, cada frecuencia de la señal sería un atleta y la pista sería un filtro paso banda donde el ancho de banda es el ancho de la pista. Los atletas tienen la misma velocidad. Si sólo hay atletas dentro de la pista, éstos llegarán al mismo tiempo, al igual que las frecuencias. Ahora, si se ponen atletas en las orillas de la pista, a estos les tomará más tiempo en llegar que a los que están dentro, lo mismo ocurre con las frecuencias que están en las orillas del ancho de banda del filtro; tendrán un retardo mayor. En este caso existe el retardo de grupo porque a algunas frecuencias (las que están cerca de las frecuencias de corte) les toma más tiempo que otras (las que viajan dentro del filtro).

Bibliografía

[1] Julius O. Smith; Jonathan S. Abel. "bark and erb bilinear transforms". *IEEE Transactions on Speech and Audio Processing*, 1999. vol 7, p. 697-708. ISSN: 1063-667.

[2] Ruben Anaya. *Notas sobre Filtros adaptables.* [Notas de clase] México: Universidad Nacional Autónoma de México, Facultad de Ingeniería., 2007?

[3] Abel Herrera Camacho. *Sistemas de Reconocimiento de Palabras Aisladas y Conectadas Usando la Transformada de Karhunen Loeve.* Director: Ralph Algazi. [Tesis de doctorado]. Universidad Nacional Autónoma de México, Facultad de Ingeniería., 2001. 151 p.

[4] Abel Herrera Camacho. *Notas de Procesamiento Digital de Voz.* [Notas de clase] México: Universidad Nacional Autónoma de México, Facultad de Ingeniería., 2008?

[5] Wai C. Chu. *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders.* USA: Wiley Interscience, 2003. 558 p. ISBN 0-471-37312-5.

[6] Aki Harma *et al.* "frequency-warped signal processing for audio applications". *108th AES convention Paris, France*, 2000. vol 48, p. 1011-1031.

[7] Bernard Widrow *et al.* "adaptive noise cancelling: Principles and applications". *Proceedings of the IEEE*, 1975. vol. 63, p. 1692-1716. ISSN: 0018-9219.

[8] Giri Shivraman; *et al. Speech Compression Using Wavelets.* Director: Dr. S. C. GADRE. [Tesis de licenciatura]. Department of electrical engineering, Veermata Jijabai Technological Institute. University of Mumbai, 2003.

[9] Matti Karjalainen *et al.* "warped filters and their audio applications". *Applications of Signal Processing to Audio and Acoustics*, 1997.

[10] Richard L Burden; Douglas Faires. *Análisis Numérico.* Óscar Palmas (trad.). 7° ed. México: Cengage Learning Editores, 2001. 839 p. ISBN: 970-686-134-3.

[11] Nadder A. Hamdy. *Applied Signal Processing: Concepts, Circuits and Systems.* [s.l.] CRC Press/Taylor & Francis, 2008. 517 p. ISBN: 9781420067026.

[12] Simon Haykin. *Adaptive Filter Theory.* 4° ed. USA: Prentice-Hall, 2002. 920 p. ISBN: 9780130901262.

[13] Alejandro Acero; Xuedong Huang; Hsiao-Wuen Hon. *Spoken Language Processing: a guide to theory, algorithm and system development.* USA: Prentice Hall PTR, 2001. 980 p. ISBN 0-13-022616-5.

[14] John H Mathews; Russel W Howell. *Complex Analysis for Mathematics and Engineering.* 5° ed. USA: Jones & Bartlett Learning, 2006. 633 p. ISBN: 978-0-7637-3748-1.

[15] Ron Hranac. *Group Delay.* [en línea]. Cisco Systems 2005. [Citado 13 de mayo de 2011]. Disponible en internet:. www.bhntampa.com/www/CenFL_SCTE/Group_delay.ppt.

[16] Satoshi IMAI. "cepstral analysis synthesis on the mel frequency scale". *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP*, 1983. vol 8, p. 93-96.

[17] Keiichi Tokuda; Takao Kobayashi. "recursive calculation of mel-cepstrum from lp coeffcients". *[s.n.]*, 1994.

[18] Satoshi IMAI; Kazuhito Koishida. "celp coding based on mel-cepstral analysis". *Acoustics, Speech, and Signal Processing, ICASSP*, 1995. vol 1, p. 33-36. ISSN: 1520-6149.

[19] Stphane Mallat. *A Wavelet Tour of Signal Processing.* 2° ed. USA: Academic Press1, 1998. 577 p. ISBN 0-12-466606-X.

[20] John G. Proakis; Dimitris Manolakis. *Tratamiento digital de señales.* Verónica Santalla del Río, José Luis Alba Castro (trad.). 4° ed. Méixico: Pearson Educación, 2007. 974 p. ISBN: 9788483223475.

[21] Alfred Mertins. *Signal Analysis: wavelets, filter banks, time-frequency transforms and applications.* USA: Wiley, 1999. 317 p. ISBN 0-471-98626-7.

[22] International Telecommunication Union. *Recommendation P.800 (08/96): SERIES P: TELEPHONE TRANSMISSION QUALITY. Methods for objective and subjective assessment of quality.* [en línea]. [Citado 13 de mayo de 2011]. Disponible en internet:. http://www.itu.int/rec/T-REC-P.800-199608-I/en.

[23] Nadeem Unuth. *Mean Opinion Score (MOS) - A Measure Of Voice Quality.* [en línea]. [Citado 13 de mayo de 2011]. Disponible en internet:. http://voip.about.com/od/voipbasics/a/MOS.htm.

[24] Lars Wanhammar. *DSP Integrated Circuits.* USA: Academic Press, 1999. 561 p. ISBN: 0-12-734530-2.

[25] Michael Weeks. *Digital Signal Processing Using MATLAB and Wavelets.* USA: Infinity Science Press LLC, 2007. 452 p. ISBN: 0-9778582-0-0.

[26] Bernard Widrow. *Adaptive Signal Processing.* USA: Prentice-Hall, 1985. 474 p. ISBN: 9780130040299.

[27] Wikipedia. *Cepstrum.* [en línea]. [Citado 13 de mayo de 2011]. Disponible en internet:. http://es.wikipedia.org/wiki/Cepstrum.

[28] Wikipedia. *Mean opinion score.* [en línea]. [Citado 13 de mayo de 2011]. Disponible en internet:. http://en.wikipedia.org/wiki/Mean_opinion_score.