Univ. Michoacana de San Nicolas de Hgo.
Facultad de Ingeniería Eléctrica
Notas de Síntesis y Reconocimiento de Voz

José Antonio Camarena Ibarrola

Agosto de 2011

Índice general

Introducción

Estas notas fueron desarrolladas como apoyo a los estudiantes que cursan la materia de Síntesis y Reconocimiento de Voz, el cual es un Tema Selecto de la carrera de Ingeniería en Computación. El programa de esta materia incluye temas de diversos libros los cuales son difíciles de conseguir y de algunos artículos científicos, las presentes notas son un compendio que se apega a los temas que indica el programa. Los ejemplos que se incluyen han sido desarrollados a mucho mayor detalle que lo que usualmente aparece en la bibliografía. Finalmente, es común que a los estudiantes de licenciatura se les dificulte la comprensión de material escrito en Inglés, estas notas son un apoyo también para aquellos alumnos con dificultades de lectura del inglés técnico del área de procesamiento digital de señales de audio y reconocimiento de patrones. Estas notas se encuentran a disposición de los estudiantes o de cualquier interesado en los temas de este curso en http://lc.fie.umich.mx/∼camarena/NotasProcDig.pdf. Se agradecen las observaciones y comentarios, favor de dirigirlos a camarena@umich.mx

Atentamente: Dr. José Antonio Camarena Ibarrola. Autor

Capítulo 1

Fundamentos de Procesamiento digital de señales

1.1. Introducción

Denotamos como $x_a(t)$ a una forma de onda continuamente variante en el tiempo mientras que representamos como $x(n)$ a una secuencia de números. Una secuencia de muestras tomadas periódicamente de una señal analógica podemos representarla como $x_a(nT)$, donde T es el periodo de muestreo.

Secuencias especiales de utilidad:

El impulso unitario:

$$\delta(n) = \begin{cases} 1 & n = 0 \\ 0 & De\ lo\ contrario \end{cases} \tag{1.1}$$

El escalón unitario:

$$u(n) = \begin{cases} 1 & n \geq 0 \\ 0 & De\ lo\ contrario \end{cases} \tag{1.2}$$

La secuencia exponencial

$$x(n) = a^n \tag{1.3}$$

si a es complejo, es decir, $a = re^{j\omega_0}$, entonces

$$x(n) = r^n e^{j\omega_0 n} = r^n(cos\omega_0 n + jsen\omega_0 n) \tag{1.4}$$

si $r = 1$ y $\omega_0 \neq 0$ entonces $x(n)$ es una senoide y si $r < 1$ y $\omega_0 \neq 0$ entonces $x(n)$ es una secuencia oscilatoria que decae exponencialmente.

El procesamiento de señales se puede definir como la transformación de una señal a una forma mas deseable en cierto sentido, tales transformaciones se pueden denotar

$$y(n) = T\big[x(n)\big] \tag{1.5}$$

La clase especial de sistemas lineales e invariantes en el tiempo (o en el espacio) (LTI) es particularmente útil, estos sistemas quedan completamente caracterizados por su respuesta al impulso unitario. Para estos sistemas la salida $y(n)$ se puede determinar como

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) = x(n) * h(n) \tag{1.6}$$

donde $h(n)$ es la respuesta del sistema al impulso unitario. La ecuación es equivalente a:

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) = h(n) * x(n) \tag{1.7}$$

Los LTIs son útiles para implementar filtros y como modelos de producción de voz

1.2. La transformada Z

1.2.1. Definición

La Transformada Z de una señal x(n) se define como:

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \tag{1.8}$$

La transformada Z inversa se define como:

$$x(n) = \frac{1}{2\pi j} \oint_C X(z) z^{n-1} dz \tag{1.9}$$

donde C es un contorno cerrado alrededor del origen del plano complejo z

La transformada Z (directa) está definida como una sumatoria, para que esta suma no produzca valores infinitos se requiere la siguiente condición de convergencia:

$$\sum_{n=-\infty}^{\infty} |x(n)||z^{-n}| < \infty \tag{1.10}$$

El conjunto de valores z donde la serie converge se denomina "Región de Convergencia" (ROC) y normalmente se puede definir como

$$R_1 < |z| < R_2 \tag{1.11}$$

Ej 1.- Sea $x(n) = \delta(n - n_0)$, entonces

$$X(z) = \sum_{n=-\infty}^{\infty} \delta(n) z^{-n} = z^{-n_0} \tag{1.12}$$

Ej 2.- Sea $x(n) = u(n) - u(n - N)$, entonces

$$X(z) = \sum_{n=-\infty}^{\infty} [u(n) - u(n - N)] z^{-n} = \sum_{n=0}^{N-1} z^{-n} \tag{1.13}$$

$$X(z) = 1 + \frac{1}{z} + ... + \frac{1}{z^{N-1}} \tag{1.14}$$

Dado que:

$$\sum_{n=0}^{N-1} r^n = \begin{cases} \frac{1-r^N}{1-r} & r \neq 1 \\ N & r = 1 \end{cases} \tag{1.15}$$

entonces:

$$X(z) = \frac{1 - z^{-N}}{1 - z^{-1}} \tag{1.16}$$

Ej 3.- Sea $x(n) = a^n u(n)$, entonces

$$X(z) = \sum_{n=-\infty}^{\infty} a^n u(n) z^{-n} = \sum_{n=0}^{\infty} a^n z^{-n} \tag{1.17}$$

$$X(z) = \sum_{n=0}^{\infty} \left(\frac{a}{z}\right)^n = \lim_{N\to\infty} \frac{1-(a/z)^N}{1-a/z} \tag{1.18}$$

Este límite existirá si se cumple $|a| < |z|$, lo cual define la región de convergencia, entonces:

$$\lim_{N\to\infty} \left(\frac{a}{z}\right)^N = 0 \tag{1.19}$$

Por lo tanto:

$$X(z) = \frac{1}{1-a/z} \tag{1.20}$$

Ej 4.-Sea $x(n) = b^n u(-n-1)$, entonces

$$X(z) = \sum_{n=-\infty}^{\infty} b^n u(-n-1) z^{-n} = \sum_{n=-\infty}^{-1} b^n z^{-n} \tag{1.21}$$

$$X(z) = \sum_{n=1}^{\infty} b^{-n} z^n = \sum_{n=1}^{\infty} \left(\frac{z}{b}\right)^n \tag{1.22}$$

$$X(z) = \sum_{n=0}^{\infty} \left(\frac{z}{b}\right)^n - 1 \tag{1.23}$$

$$X(z) = \lim_{N\to\infty} \frac{1-(z/b)^N}{1-z/b} - 1 \tag{1.24}$$

Este límite existirá si se cumple $|z| < |b|$, lo cual define la región de convergencia, entonces:

$$\lim_{N\to\infty} \left(\frac{z}{b}\right)^N = 0 \tag{1.25}$$

Por lo tanto:

$$X(z) = \frac{1}{1-z/b} - 1 \tag{1.26}$$

Tabla 1.1: Propiedades de la transformada Z

Linealidad	$ax_1(n) + bx_2(n)$	$aX_1(z) + bX_2(z)$
desplazamiento	$x(n + n_0)$	$z^{n_0}X(z)$
Ponderación por exponencial	$a^n x(n)$	$X(a^{-1}z)$
Ponderación lineal	$nx(n)$	$-z\frac{dX(z)}{dz}$
Inversión de tiempo	$x(-n)$	$X(z^{-1})$
Convolución	$x(n) * h(n)$	$X(z)H(z)$
Producto de secuencias	$x(n)w(n)$	$\frac{1}{2\pi j}\oint_C X(v)W(z/v)v^{-1}dv$

Como en los ejemplos 3 y 4, es típico para secuencias con valores diferentes de cero para $n < 0$ tener como región de convergencia $|z| < |R|$ y es típico para secuencias con valores diferentes de cero para $n > 0$ tener como región de convergencia $|z| > |R|$. Para secuencias con valores diferentes de cero tanto para $n > 0$ como para $n < 0$ tendremos una región de convergencia de tipo $|R_1| < |z| < |R_2|$

1.2.2. Propiedades de la Transformada Z

La transformada Z tiene varias propiedades, algunas de las mas importantes se muestran en la Tabla 1.1.

Demostremos la propiedad de desplazamiento

Sea

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \tag{1.27}$$

$$\sum_{n=-\infty}^{\infty} x(n-n_0)z^{-n} = \sum_{m=-\infty}^{\infty} x(m)z^{-m-n_0} = z^{-n_0}\sum_{m=-\infty}^{\infty} x(m)z^{-m} = z^{-n_0}X(z) \tag{1.28}$$

donde $m = n - n_0$

1.3. La Transformada de Fourier

1.3.1. Definición a partir de la Transformada Z

La transformada de Fourier es un caso especial de la transformada Z en el cual $z = e^{j\omega}$. Entonces:

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \tag{1.29}$$

$$x(n) = \int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}d\omega \tag{1.30}$$

Observe que la integral de contorno se convirtió en una integral de $-\pi$ a π que realmente está integrando los valores alrededor del circulo trigonométrico unitario.

Basta sustituir $e^{j\omega}$ por z en la Tabla 1.1 para obtener un conjunto de propiedades de la Transformada de Fourier.

1.3.2. La Transformada Discreta de Fourier

En el caso de que una secuencia $x(n)$ sea periódica (con periodo N), es decir $x(n) = x(n+N)$ entonces podemos usar una serie de Fourier para representarla como una suma de senoides. La Transformada Discreta de Fourier (DFT) se define como:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N} \tag{1.31}$$

Mientras que la Transformada Discreta Inversa de Fourier (IDTF) se define como:

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{j2\pi kn/N} \tag{1.32}$$

La DFT ($X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi kn/N}$) tiene un kernel de tamaño proporcional al de la señal

$$\left[\begin{array}{c} X(0) \\ \hline X(1) \\ \hline X(2) \\ \hline \vdots \\ \hline X(N) \end{array}\right] = \left[\begin{array}{c|c|c|c|c} 1 & 1 & 1 & \cdots & 1 \\ \hline 1 & W_N^1 & W_N^2 & \cdots & W_N^{N-1} \\ \hline 1 & W_N^2 & W_N^4 & \cdots & W_N^{2(N-1)} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline 1 & W_N^{N-1} & W_N^{2(N-1)} & \cdots & W_N^{(N-1)^2} \end{array}\right] \left[\begin{array}{c} x(0) \\ \hline x(1) \\ \hline x(2) \\ \hline \vdots \\ \hline x(N) \end{array}\right]$$

donde $W_N = e^{-j2\pi/N}$

Por tanto, la complejidad para determinarla directamente es $O(N^2)$. La transformada discreta de Fourier puede de hecho determinarse con $O(Nlog_2N)$ operaciones con un algoritmo conocido como la Transformada Rápida de Fourier (FFT). La diferencia es inmensa para cierta N equivale a poderla determinar en 30 segundos en lugar de 2 semanas de tiempo de CPU. La existencia de un algoritmo FFT se conoció hasta mediados de los años sesentas a partir del trabajo de J.W. Cooley y J.W. Tukey [1]. Sabemos de varios métodos de cálculo eficiente de la DFT (ver [2]), el de Danielson and Lanczos in 1942, proveé una de las mas claras derivaciones del algoritmo. Danielson y Lanczos mostraron que la DFT de una secuencia de longitud N podía escribirse como la suma de dos DFTs cada una de longitud $N/2$, una de ellas se conforma de los valores ubicados en posiciones pares y la otra de los valores ubicados en posiciones impares.

$$X(k) = \sum_{n=0}^{N-1} e^{-j2\pi kn/N} x(n) \tag{1.33}$$

$$X(k) = \sum_{n=0}^{N/2-1} e^{-j2\pi k(2n)/N} x(2n) + \sum_{n=0}^{N/2-1} e^{-j2\pi k(2n+1)/N} x(2n+1) \tag{1.34}$$

$$X(k) = \sum_{n=0}^{N/2-1} e^{-j2\pi kn/(N/2)} x(2n) + W^k \sum_{n=0}^{N/2-1} e^{-j2\pi kn/(N/2)} x(2n+1) \tag{1.35}$$

$$X(k) = X^e(k) + W_N^k X^o(k) \tag{1.36}$$

$X^e(k)$ denota el k-ésimo componente de la DFT de la señal de longitud $N/2$ formada a partir de los componentes pares de la señal original $X^o(k)$ es

el k-ésimo componente de la DFT de la señal de longitud $N/2$ formada con los componentes impares de la señal original.

El lema de Danielson-Lanczos puede usarse recursívamente. Una vez que se ha reducido el problema de calcular la transformada de una señal de longitud N al problema de calcular dos transformadas $X^e(k)$ y $X^o(k)$ cada una de longitud $N/2$ podemos hacer la misma reducción con $X^e(k)$ y $X^o(k)$ a calcular la transformada de sus $N/4$ datos pares y sus $N/4$ datos impares, podemos definir $X^{ee}(k)$ y $X^{eo}(k)$ como las transformadas discretas de los valores ubicados en posiciones par-par y par-impar en las sucesivas subdivisiones de la señal. Es conveniente que N sea una potencia de 2, si no es así se pueden concatenar ceros a la señal hasta que la longitud de esta sea una potencia de dos. Una vez salvada esta restricción, es evidente que podemos continuar aplicando el lema de Danielson-Lanczos hasta que la transformada sea de longitud 1. La transformada de Fourier de longitud uno es simplemente la operación identidad. En otras palabras, para cada patron de $log_2 N$ e's y o's hay una transformada de un solo valor que es simplemente uno de los valores de la señal de entrada x(n).

$$X^{eoeeoeo...oee}(k) = x(n) \tag{1.37}$$

para alguna n

Lo que falta es averiguar cual n corresponde a cada patron de e's y o's. La respuesta es: Al invertir el patron de e's y o's y después hacer $e = 0$ y $o = 1$ se obtendrá en binario el valor de n, esto se debe a que las sucesivas subdivisiones de los datos en localidades pares e impares son en realidad pruebas del bit menos significativo de n. La idea de la inversión de bits junto con el lema de Danielson-Lanczos sirve para hacer un algoritmo práctico de FFT: Suponga que se toma el vector original de datos x y lo rearreglamos en el order de bits invertidos como en la Figura 1.1.

Entonces, los valores ya no estarán en el orden de n sino en el que resulta de invertir n a nivel de bits. Los valores del arreglo se interpretan ahora como transformadas de Fourier de longitud uno. Combinamos ahora parejas de valores adyacentes para obtener transformadas de "señales" de longitud dos, luego combinamos parejas de parejas para obtener transformadas de "señales" de longitud 4 y así sucesívamente hasta que la primera mitad del arreglo se combine con la segunda mitad del arreglo para obtener la transformada final. Cada combinación toma del oren de N operaciones y como hay $log_2 N$ combinaciones, entonces el algortimo completo es del orden $N log_2 N$.

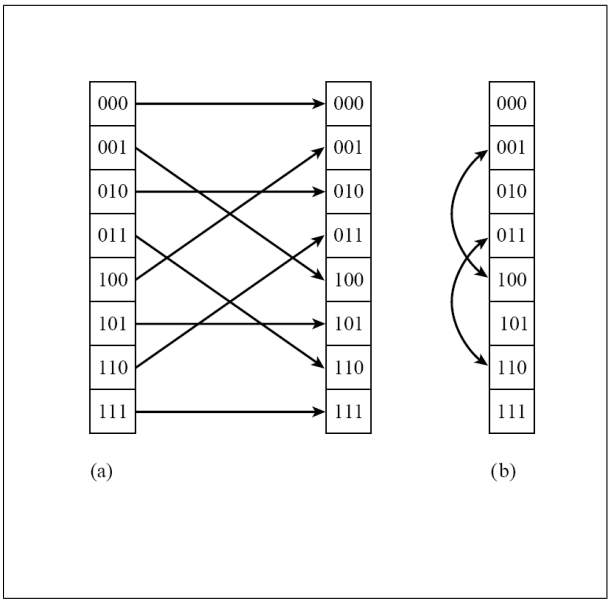

Figura 1.1: Reordenamiento de un arreglo por inversión de bits. (a) en dos arreglos. (b) en el mismo arreglo

En la implementación de este algoritmo se debe cuidar de no llamar a la función seno o a la coseno mas veces de las necesarias, si no se tiene ese cuidado se puede invocar a la función seno o a la coseno con el mismo ángulo muchas veces.

1.3.3. Propiedades y Limitaciones

La Transformada discreta de Fourier requiere que la señal sea estacionaria, es decir, que los contenidos de frecuencia de la señal no estén cambiando en el tiempo, cuando este requerimiento no se cumple, la DFT asumirá que los diversos contenidos de frecuencia encontrados a lo largo del tiempo estaban presentes de principio a fin. La transformada Continua de Fourier no tiene tal limitación, no debemos ver a la Transformada Discreta de Fourier como una discretización de la Transformada Continua de Fourier, para entender esto, considere el siguiente ejemplo, sea:

$$x(t) = sin(\frac{t^2}{2}) \tag{1.38}$$

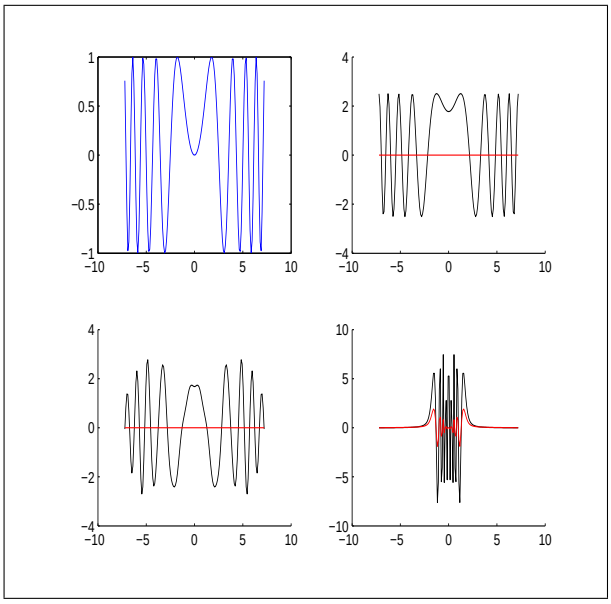

Figura 1.2: Ejemplo de una señal no estacionaria, su transformada Continua de Fourier y su Transformada Discreta de Fourier

Su Transformada continua de Fourier se determina mediante una integral que se puede resolver mediante técnicas de integración, aunque existen tablas de Transformadas de Fourier que nos pueden ahorrar el esfuerzo.

$$X(j\omega) = \int_{-\infty}^{\infty} sin(\frac{t^2}{2})e^{-j\omega t}dt \tag{1.39}$$

$$= \pi[cos(\frac{\omega^2}{2}) + sin(\frac{\omega^2}{2})] \tag{1.40}$$

En la Figura 1.2 se muestra la señal no estacionaria $x(t) = sin(\frac{t^2}{2})$, su Transformada Continua, una Discretización de la transformada continua y la Transformada Discreta de Fourier, la línea roja representa la parte imaginaria y la línea negra es la parte real, en este ejemplo, la parte imaginaria de la Transformada Continua de Fourier es cero como se aprecia en la gráfica de la Transformada Continua de Fourier así como en la Discretización, sin embargo, la Transformada Discreta de Fourier falla al entregar una parte imaginaria diferente de cero.

1.3.4. Diseño de un filtro usando la DFT y la DFT inversa

Podemos implementar cualquier tipo de filtro pasa-banda mediante una técnica simple que consiste en los siguientes pasos:

1. Aplicar la Transformada Discreta de Fourier a la señal de entrada
2. Dados los valores de frecuencia que delimitan la banda, determinar cuales coeficientes espectrales caen fuera de esta. Para ello conviene recordar la fórmula para determinar la frecuencia del k-ésimo coeficiente espectral:

$$f(Hz) = \frac{(k)(f_m)}{N} \tag{1.41}$$

   donde f_m es la frecuencia de muestreo y N es el número de muestras de la señal.
3. Hacer cero a los coeficientes espectrales que caen fuera de la banda (Tanto a la parte real como a la parte imaginaria), hacer lo mismo con las frecuencias negativas.
4. Aplicar Transformada de Fourier Inversa
5. La señal resultante es la señal filtrada

Para mejorar la calidad del filtro, en lugar de simplemente borrar los coeficientes espectrales que caen fuera de la banda se puede multiplicar la secuencia de coeficientes espectrales por una gaussiana cuyo centro coincida con el centro de la banda de paso. El ancho de la gaussiana debe coincidir con el ancho de la banda de paso. De esta manera no solo los coeficientes que están lejos de la gaussiana terminarán valiendo prácticamente cero sino que dentro de la banda los coeficientes que estén mas cerca del centro de la banda tendrán mayor peso que los que estén cerca de los extremos. Después se aplica la transformada inversa de la misma manera. Tome en cuenta que en realidad son dos gaussianas las que se deben aplicar, una para las frecuencias positivas y otra para las frecuencias negativas, y hacerlo tanto para las partes reales como para las partes imaginarias.

En forma similar se pueden implementar filtros pasa-bajas, pasa-altas o de rechazo de banda.

1.4. Filtros Digitales

Un filtro digital es un sistema lineal invariante en el tiempo, para tales sistemas la salida $y(n)$ se puede determinar mediante una convolución de la entrada $x(n)$ con la respuesta al impulso unitario $h(n)$, en el dominio z esto es:

$$Y(z) = H(z)X(z) \tag{1.42}$$

$H(z)$, la llamada función sistema es la transformada Z de la respuesta del sistema al impulso unitario $h(n)$, mientras que a $H(e^{j\omega})$ (la transformada de Fourier de $h(n)$) se le conoce como su respuesta a la frecuencia.

1.4.1. Causalidad

Un sistema Causal es aquel en el cual no existe una salida diferente de cero cuando a la entrada todavía no se aplica nada, es decir, si no hay excitación, no hay respuesta en un sistema causal. Todos los sistemas prácticos (reales) son causales. Formalmente, en un sistema causal $h(n) = 0$ para todo $n < 0$.

1.4.2. Estabilidad

Un sistema estable es aquel en el que a una entrada acotada produce una salida acotada.

1.4.3. Sistemas lineales e invariantes en el tiempo

Todos los sistemas lineales invariantes en el tiempo que son de interés para la implementación de filtros digitales satisfacen una ecuación de diferencias de la forma:

$$y(n) - \sum_{k=1}^{N} a_k y(n-k) = \sum_{r=0}^{M} b_r x(n-r) \tag{1.43}$$

Aplicando transformada Z a ambos lados obtenemos:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{r=0}^{M} b_r z^{-r}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{1.44}$$

Como se aprecia, la función sistema se puede obtener simplemente extrayendo los coeficientes de la ecuación de diferencias. La función sistema conviene para determinar sus polos y ceros expresarla en la forma siguiente:

$$H(z) = \frac{A\Pi_{r=0}^{M}(1 - c_r z^{-1})}{\Pi_{k=1}^{N}(1 - d_k z^{-1})} \tag{1.45}$$

Existen dos clases de sistemas lineales invariantes en el tiempo, es decir, aquellos cuya respuesta al impulso tiene una duración finita (FIR) y aquellos cuya respuesta al impulso tiene una duración infinita (IIR)

1.4.4. Filtros FIR

Si todos los coeficientes a_k de la ecuación (1.43) valen cero la ecuación de diferencias se convierte en:

$$y(n) = \sum_{r=0}^{M} b_r x(n - r) \tag{1.46}$$

La respuesta al impulso es :

$$h(n) = \begin{cases} b_n & 0 \leq n \leq M \\ 0 & de\ otro\ modo \end{cases} \tag{1.47}$$

$H(z)$ es un polinomio en z^{-1} y por tanto no tiene polos, solo ceros. Los sistemas FIR tienen la propiedad de fase lineal exacta.

Si $h(n)$ satisface la relación:

$$h(n) = \pm h(M - n) \tag{1.48}$$

entonces $H(e^{j\omega})$ tiene la forma $A(e^{j\omega})e^{-j\omega(M/2)}$

La propiedad de fase lineal consiste en que la respuesta en la fase de un filtro depende linealmente de la frecuencia, esta propiedad es muy importante en aplicaciones de voz donde el alineamiento en tiempo es esencial para simplificar el problema de la precisión.

La desventaja de los filtros FIR es que no es fácil lograr que corten agudamente a la frecuencia de corte comparados con los filtros IIR.

Una ecuación de diferencias se puede representar mediante un diagrama de bloques, la simbología utilizada en los mismos se resume en la Fig. 1.3

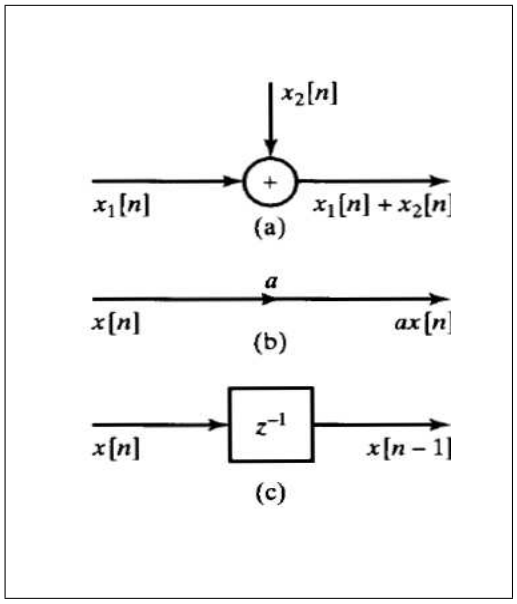

Figura 1.3: Simbología utilizada en diagramas de bloques de filtros digitales

La ecuación de diferencias (1.43) se puede representar entonces mediante el diagrama de bloques mostrado en la Figura 1.4.

1.4.5. Diseño de un filtro FIR

Solo los filtros FIR pueden tener fase lineal. Un filtro FIR de longitud M tiene respuesta a la frecuencia:

$$H(\omega) = \sum_{k=0}^{M-1} b_k e^{-j\omega k} \tag{1.49}$$

donde los coeficientes b_k son los valores que se obtienen a la salida del sistema cuando se aplica un impulso unitario.

1.4.5.1. Determinación de los coeficientes del filtro mediante la técnica del muestreo de la frecuencia

La condición de fase lineal se logra imponiendo restricciones de simetría a la respuesta del filtro al impulso unitario, es decir:

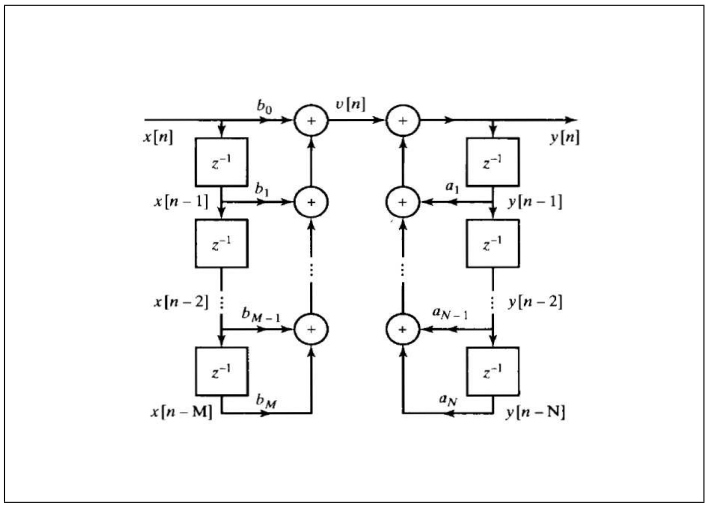

Figura 1.4: La ecuación de diferencias (1.43) representada mediante un diagrama de bloques

$$h(n) = h(M-1-n) \tag{1.50}$$

Seleccionando valores equiespaciados de frecuencia en el rango $0 \leq \omega \leq \pi$:

$$\omega_k = \frac{2\pi k}{M} \quad \forall k = \begin{cases} 0, 1, ..., \frac{M-1}{2} & M \ \ impar \\ 0, 1, ..., \frac{M}{2} - 1 & M \ \ par \end{cases} \tag{1.51}$$

Para diseñar el filtro FIR simétrico resolvemos el sistema de ecuaciones siguiente (suponiendo que M es impar):

$$\sum_{n=0}^{(M-1)/2} a_{kn} h(n) = H_r(\omega_k) \quad \forall k = 0, 1, ..., \frac{M-1}{2} \tag{1.52}$$

donde:

$$a_{kn} = 2cos\omega_k \Big(\frac{M-1}{2} - n \Big) \tag{1.53}$$

$$a_{kn} = 1 \quad n = \frac{M-1}{2} \ \ \forall k \tag{1.54}$$

en caso de que M sea par el sistema de ecuaciones a resolver es:

$$\sum_{n=0}^{(M/2)-1} a_{kn}h(n) = H_r(\omega_k) \quad \forall k = 0, 1, ..., \frac{M}{2} - 1 \tag{1.55}$$

Ejemplo: Determinar los coeficientes (respuesta al impulso unitario h(n)) de un filtro FIR con fase lineal y longitud $M = 4$ cuya respuesta a la frecuencia para $\omega = 0$ y $\omega = \pi/2$ es:

$$H_r(0) = 1, \quad H_r(\pi/2) = 1/2 \tag{1.56}$$

Solución: Las restricciones de simetría (para que el filtro tenga fase lineal) dicen que $h(0) = h(3)$ y que $h(1) = h(2)$ por lo que solo hay que determinar dos coeficientes, estos se obtienen resolviendo el sistema de ecuaciones:

$$\begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix} \begin{bmatrix} h(0) \\ h(1) \end{bmatrix} = \begin{bmatrix} H_r(0) \\ H_r(\pi/2) \end{bmatrix} \tag{1.57}$$

en este caso:

$$\begin{bmatrix} 2 & 2 \\ -\sqrt{2} & \sqrt{2} \end{bmatrix} \begin{bmatrix} h(0) \\ h(1) \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \end{bmatrix} \tag{1.58}$$

de donde: $h(0) = h(3) = 0{,}0732232$ y $h(1) = h(2) = 0{,}4267766$

El siguiente código de MATLAB determina los coeficientes de un filtro FIR con fase lineal por el método de muestreo de la frecuencia

```
function h=CoefsFIR(H,M)
%M es la longitud del filtro
%H es la secuencia de valores equiespaciados de 0 a pi de la funcion sistema d
%ej H=[1 1 1 1 0.4 0 0 0]' (ojo transpuesta) para M=15
if mod(M,2)==0
    a=zeros(M/2,M/2);
    for k=0:M/2-1
       for n=0:M/2-1
          a(k+1,n+1)=2*cos((2*pi*k/M)*((M-1)/2-n));
      end
  end
else
    a=zeros((M-1)/2,(M-1)/2);
```

```
    for k=0:(M-1)/2
       for n=0:(M-1)/2
          a(k+1,n+1)=2*cos((2*pi*k/M)*((M-1)/2-n));
      end
      a(k+1,(M-1)/2+1)=1;
  end
end h=inv(a)*H;
```

Por ejemplo para diseñar un filtro con $M = 15$ con la siguiente especificación:

$$H_r\Big(\frac{2\pi k}{15}\Big) = \begin{cases} 1 & k = 0, 1, 2, 3 \\ 0{,}4 & k = 4 \\ 0 & k = 5, 6, 7 \end{cases} \tag{1.59}$$

Invocaríamos la función de la siguiente manera:

```
>> CoefsFIR([1 1 1 1 0.4 0 0 0]',15)

ans =

   -0.0141
   -0.0019
    0.0400
    0.0122
   -0.0914
   -0.0181
    0.3133
    0.5200
```

1.5. Filtros IIR (Respuesta de duración infinita al impulso)

Cuando en la ecuación de diferencias (1.43) existen tanto coeficientes $a_k \neq 0$ como $b_k \neq 0$ la función sistema tendrá tanto polos como ceros (y no solo ceros como es el caso de los sistemas FIR). La ecuación de diferencias (1.43) nos indica que los valores en la salida del sistema se pueden determinar a

partir de los valores presentes y pasados de la entrada. Es posible implementar sistemas IIR de manera que el cálculo de la salida sea mas eficiente, es decir, se realice con menos cálculos, especialmente cuando se implementan filtros de orden alto los cuales cortan de manera mas aguda al pasar la frecuencia de corte.

Existe una variedad de métodos para diseñar filtros. Para filtros cuya función es la de ser selectivos a la frecuencia (filtros pasabajas, pasaltas, etc.) los métodos de diseño están generalmente basados en procedimientos clásicos como:

1. Diseño de Butterworth
2. Diseño de Bessel
3. Diseño de Chebyshev
4. Diseños elípticos

La diferencia mas importante entre los filtros FIR y los IIR es que los filtros IIR no se pueden diseñar para que tengan respuesta de fase lineal, en cambio, los filtros IIR son mas eficientes en el sentido de lograr un corte agudo al sobrepasar la frecuencia de corte.

1.5.1. Implementación de filtros IIR con requerimientos mínimos de almacenamiento

La implementación de filtros IIR es flexible, la implementación directa es la de la Figura 1.4. La ecuación de diferencias (1.43) se puede transformar a formas equivalentes, así entonces:

$$w(n) = \sum_{k=1}^{N} a_k w(n-k) + x(n) \tag{1.60}$$

$$y(n) = \sum_{r=0}^{M} b_r w(n-r) \tag{1.61}$$

Las ecuaciones anteriores se implementan como se indica en la Fig. 1.5 con un significativo ahorro de la memoria requerida para almacenar la secuencia de valores retrasados

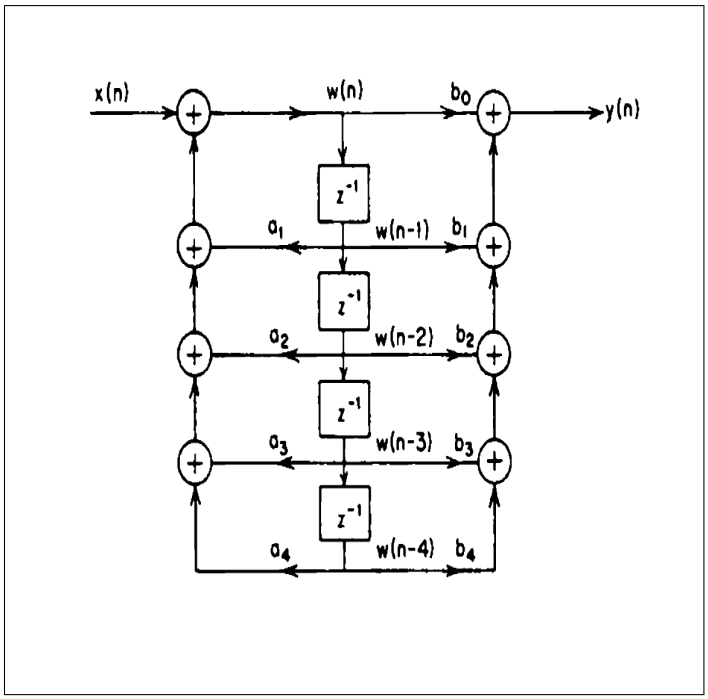

Figura 1.5: Implementación de un sistema IIR de manera que se ahorra memoria

1.5.2. Diseño de filtros IIR mediante transformación Z apareada

Este método consiste en mapear los polos y los ceros de la función de transferencia en el dominio de Laplace s al plano complejo z. más precisamente a algún lugar dentro del círculo trigonométrico unitario para efecto de garantizar la estabilidad del filtro. Esta idea es explotada por tres métodos de diseño de filtros IIR, a saber, la transformación directa, la transformación Z apareada y la transformación bilineal. En el caso de la transformada Z apareada, por cada frecuencia crítica (polo o cero) a en el dominio de Laplace, existirá una frecuencia crítica correspondiente en el plano complejo Z ubicado en:

$$e^{aT_s} \tag{1.62}$$

donde T_s es el periodo de muestreo, es decir el recíproco de la frecuencia de muestreo f_s.

La transformación que este método implementa se puede expresar como:

$$s - a \rightarrow 1 - e^{aT_s} z^{-1} \tag{1.63}$$

Ejemplo: Diseñe un filtro digital IIR que simule al filtro analógico cuya función de transferencia es:

$$H_s(s) = \frac{s}{s^2 + 400s + 2 \times 10^5} \tag{1.64}$$

para una frecuencia de muestreo $f_s = 1KHz$.

La función de transferencia tiene un cero en $s = 0$ y polos en $s = -200 \pm j400$. Entonces, en el dominio Z, tendremos un cero en $z = 1$ y polos en:

$$z = e^{(-200 \pm j400)T_s} = e^{-0,2 \pm j0,4} = 0,7541 \pm j0,3188 \tag{1.65}$$

Por lo que la función de transferencia en el dominio Z es:

$$H_z(z) = \frac{1 - z^{-1}}{(1 - (0,7541 + j0,3188)z^{-1})(1 - (0,7541 - j0,3188)z^{-1})} \tag{1.66}$$

de donde

$$H_z(z) = \frac{1 - z^{-1}}{1 - 1,509z^{-1} + 0,6708z^{-2}} \tag{1.67}$$

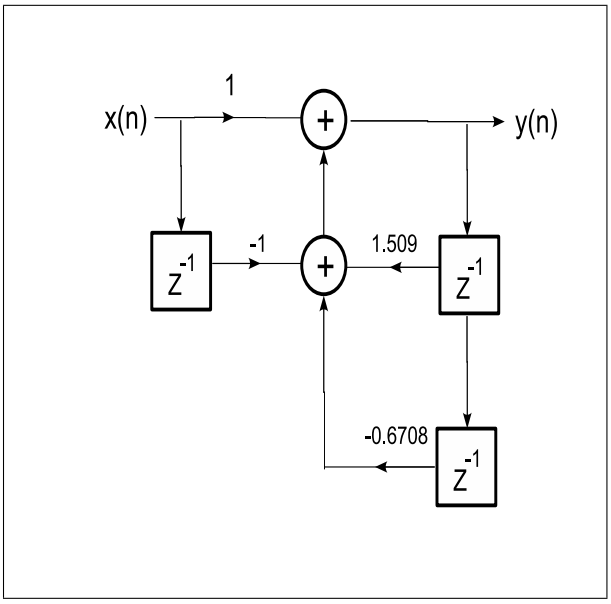

Figura 1.6: Filtro IIR diseñado mediante mapeo de frecuencias críticas

La ecuación de diferencias sería entonces:

$$y(n) = x(n) - x(n-1) + 1{,}509y(n-1) - 0{,}6708y(n-2) \qquad (1.68)$$

El diagrama de bloques correspondiente sería el mostrado en la Figura 1.6.

1.5.2.1. Uso de MATLAB

Para diseñar filtros IIR mediante el método "invariante al impulso" (Una variante del método aquí explicado), MATLAB cuenta con la función:

[BZ, AZ]= IMPINVAR(B,A,Fs)

donde: BZ y AZ son los coeficientes del polinomio del numerador y del denominador respectivamente de la función sistema (Función de transferencia en el dominio Z) de un filtro digital IIR cuya respuesta al impulso es idéntica a la respuesta al impulso de un filtro analógico cuya función de transferencia en el dominio de Laplace es una fracción donde B y A son los coeficientes

del polinomio de numerador y del denominador respectivamente, Fs es la frecuencia de muestreo. Para el ejemplo anterior usaríamos la función de la siguiente manera:

```
>> [BZ,AZ]=impinvar([1 0],[1 400 200000],1000)

BZ =

    0.0010   -0.0009


AZ =

    1.0000   -1.5082    0.6703
```

1.5.3. Filtros Butterworth

Como vimos, podemos diseñar un filtro discreto IIR transformando un filtro analógico que cumpla ciertas especificaciones en uno digital, por esta razón es necesario conocer al menos una técnica de diseño de filtros analógicos y el filtro Butterworth es probablemente el filtro analógico más sencillo de diseñar.

El filtro de Butterworth, Se diseña para tener a respuesta de frecuencia tan plana como matemáticamente sea posible en la banda de paso. Otro nombre para ellos es filtros "de magnitud máximamente plana". Tienen una caída suave en la región de transición, la rapidez de la caída en la región de transición aumenta con el orden del filtro. Normalmente se usa como filtro antialias para señales analógicas que van a ser muestreadas. El filtro de Butterworth fue presentado por el Ingeniero Británico Stephen Butterworth.

Visto en un diagrama de Bode con escala logarítmica, la respuesta decae linealmente desde la frecuencia de corte hacia menos infinito. Para un filtro de primer orden son -20 dB por década (aprox. -6dB por octava).

El filtro de Butterworth es el único filtro que mantiene su forma para órdenes mayores (sólo con una caída de mayor pendiente a partir de la frecuencia de corte).

Este tipo de filtros necesita un mayor orden para los mismos requerimientos en comparación con otros, como los de Chebyshev o el elíptico.

Si llamamos H a la respuesta en frecuencia, se debe cumplir que las $2N-1$ primeras derivadas de $|H(\Omega)|^2$ sean cero para $\Omega = 0$ y $\Omega = \infty$

Únicamente posee polos y la función de transferencia es:

$$|H(\Omega)|^2 = \frac{1}{1 + (\Omega/\Omega_c)^{2N}} \tag{1.69}$$

donde N es el orden del filtro, Ω_c es la frecuencia de corte (en la que la respuesta cae 3 dB por debajo de la banda pasante) y Ω es la frecuencia analógica compleja ($\Omega = j\omega$).

La función de transferencia de un filtro pasabajas de primer orden es:

$$\frac{V_o}{V_i} = \frac{G}{\omega_0 s + 1} \tag{1.70}$$

La función de transferencia de un filtro pasa bajas de segundo orden es:

$$\frac{V_o}{V_i} = \frac{G}{s^2 + \frac{\omega_0}{Q} s + \omega_0^2} \tag{1.71}$$

Los filtros Butterworth solo tienen polos y estos se distribuyen uniformemente a lo largo del perímetro del círculo trigonométrico unitario. Para que el filtro sea estable se utilizan solo los polos en el semiplano izquierdo.

1.6. Muestreo

1.6.1. La regla de Nyquist

Si una señal $x_a(t)$ tiene una transformada de Fourier $X_a(j\omega)$ tal que $X_a(j\omega) = 0$ para $\omega \geq 2\pi F_N$, entonces $x_a(t)$ puede ser reconstruida a partir de muestras equiespaciadas $x_a(nT)$ si y solo si $2F_N < 1/T$. A la frecuencia F_N le llamamos la frecuencia de Nyquist.

El teorema del muestreo tiene la implicación de que no tiene sentido muestrear a una frecuencia mayor que la frecuencia de Nyquist (no se puede pedir una reconstrucción mejor que la reconstrucción perfecta) y entonces se considera un desperdicio de espacio el tomar mas muestras de las estrictamente necesarias.

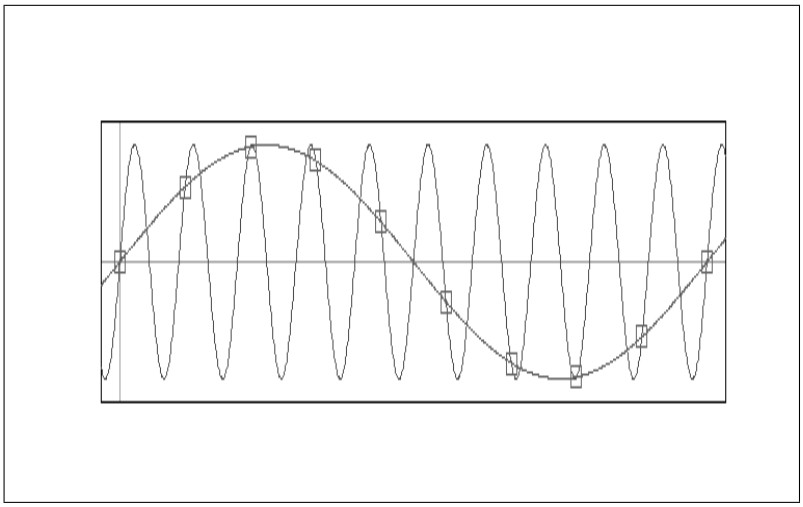

Figura 1.7: Efecto denominado Aliasing provocado por muestrear a una frecuencia inferior a la de Nyquist

1.6.1.1. El efecto Aliasing

La regla de Nyquist nos dice que se debe muestrear con una frecuencia de al menos el doble de la frecuencia mayor que aún es de interés, esta regla nos previene que en caso de no cumplir con esto, ocurriría el efecto "Aliasing" que consiste en que una señal de frecuencia alta es confundido con su "alias" que es una señal de frecuencia baja al muestrear la señal a una frecuencia de menos del doble, ver la Figura 1.7.

1.6.1.2. Filtro antialiasing

Una vez que se ha decidido muestrear a determinada frecuencia de muestreo, no podremos ver los componentes de la señal cuya frecuencia sea mayor de la mitad de de frecuencia de muestreo, estos sería confundidos con sus alias, por lo tanto conviene antes de digitalizar la señal mediante un convertidor analógico-digital (ADC) pasar la señal por un filtro "anti-aliasing" que es un filtro pasabajas con una frecuencia de corte de la mitad de la frecuencia de muestreo, este filtro casi siempre viene incluido en las tarjetas digitalizadoras como las tarjetas de audio.

1.7. Submuestreo crítico

Submuestrear una señal es la operación de desechar muestras de la señal, esto se hace cuando se considera que se cumple de sobra con la regla de Nyquist y que las muestras adicionales son en realidad redundantes, en un diagrama representamos el submuestreo con una flecha hacia abajo ($\downarrow$). Si el submuestreo consiste en desechar todas las muestras que ocupan posiciones pares y conservar aquellas que ocupan posiciones pares, entonces lo denotamos en el diagrama como $2\downarrow$.

El submuestreo crítico es mayor submuestreo que se puede hacer de modo que con la señal submuestreada todavía se puede hacer la *reconstrucción perfecta*. La reconstrucción perfecta es aquella en la que la señal reconstruida es prácticamente una copia de la señal original donde no hay mas distorsión que un desplazamiento en el tiempo y/o una escalamiento en la amplitud de la señal.

Capítulo 2

Modelos digitales de la señal de voz y Teoría acústica

La señal de voz se puede considerar como una secuencia de sonidos producidos por el tracto vocal-nasal, al estudio de las reglas para arreglar estos sonidos de manera que esta secuencia de sonidos se puedan interpretar como información se le conoce como "Lingüística" y a la disciplina que se dedica a la clasificación de estos sonidos y como se producen en el tracto vocal-nasal se le llama "fonética". Necesitamos conocer un poco acerca de la clasificación de estos sonidos para poder hacer labores de reconocimiento, compresión o síntesis de voz.

2.1. Anatomía y fisiología de la Producción de voz

La Figura 2.2 muestra el diagrama esquemático del tracto vocal-nasal humano. Este diagrama resalta las características físicas esenciales de la anatomía humana que participan en las etapas finales del proceso de producción de la voz. Se muestra el tracto vocal como un tubo cuya área de sección transversal no es uniforme y que está acotada en un extremo por las cuerdas vocales y en el otro extremo por los labios. Este tubo sirve como un sistema de transmisión acústica. Para sonidos nasales una ramificación denominada "tracto nasal" se conecta a al flujo principal mediante una compuerta a la que llamamos "velo". La forma del tubo, básicamente el área de sección a lo largo del eje longitudinal varía en el tiempo debido a los movimientos de los

labios, la mandíbula, la lengua y el velo. Aunque el tracto vocal humano no está en línea recta como en la Figura 2.2, este modelo es una aproximación razonable.

2.2. Fonética: sonidos vocalizados y no-vocalizados

Los sonidos que produce el tracto vocal-nasal pueden clasificarse de diferentes maneras. Los sonidos vocalizados son producidos cuando el tracto vocal es excitado por pulsos de presión de aire que resultan de aperturas y cierres cuasi-periódicos del orificio glotal (entre las cuerdas vocales). Los sonidos no-vocalizados se producen forzando un flujo de aire turbulento que actúa como una excitación de ruido aleatorio inyectado al tracto vocal. Una tercera forma de producir sonido ocurre cuando el tracto vocal se cierra parcialmente provocando flujo turbulento pero al mismo tiempo permitiendo flujo cuasi-período debido a la vibración de las cuerdas vocales, a los sonidos producidos de esta manera les llamamos fricativos vocalizados. Finalmente, plosivos y africados se forman interrumpiendo momentáneamente el flujo de aire, permitiendo que aumente la presión de aire y liberando repentinamente esa presión. Para todas estas formas de excitación, el tracto vocal actúa como una línea de transmisión acústica con ciertas resonancias dependientes de la forma del tracto que enfatizan algunas frecuencias de la excitación.

2.3. El concepto de Formantes

La señal de voz varía a razón de 10 fonemas por segundo pero las variaciones de la forma de onda suceden a un régimen mucho mayor, es decir, los cambios en la configuración del tracto vocal ocurren lentamente en comparación con las variaciones detalladas de la señal de voz. Los sonidos creados en el tracto vocal van tomando forma en el dominio de la frecuencia dependiendo de la respuesta a la frecuencia que tenga el tracto vocal. Las frecuencias de resonancia de una particular configuración de las articulaciones son los instrumentos que forman el sonido para darle forma a un fonema dado, por eso es que a estas frecuencias de resonancia se les llama “frecuencias formantes del sonido” o simplemente “formantes”. En resumen, la estructura fina de la forma de onda es creada por las fuentes del sonido en el tracto vocal y las resonancias del tracto vocal le dan forma a estas fuentes de sonido para

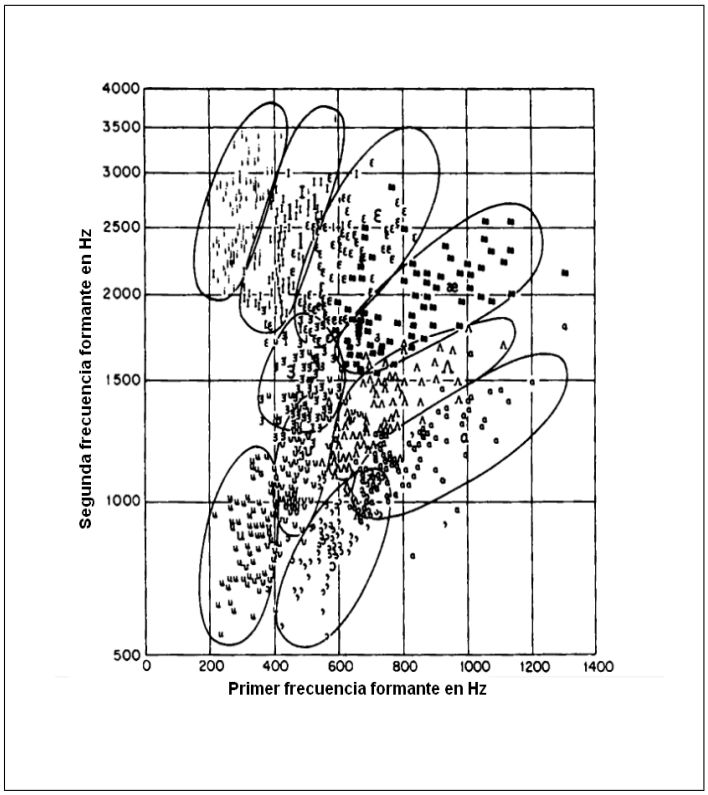

Figura 2.1: Los primeros dos formantes se pueden utilizar para clasificar a los sonidos vocalizados

convertirlas en fonemas.

2.3.1. Mapa de formantes para los sonidos vocalizados mas comunes

En la Figura 2.1 se muestran para varios sonidos vocalizados como se distribuyen en un espacio de características bidimensional donde el eje horizontal es el primer formante y el eje vertical es el segundo formante.

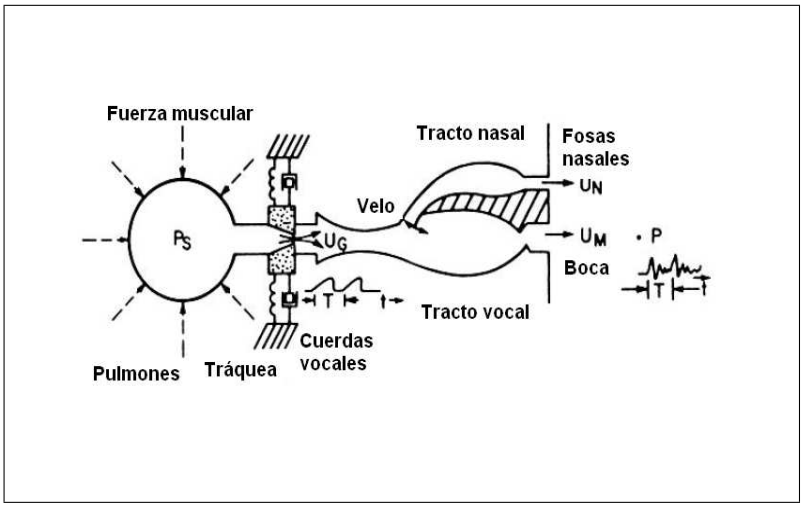

Figura 2.2: Diagrama esquemático del Tracto Vocal-nasal humano

2.3.2. Técnica de síntesis de sonidos vocalizados usando los primeros tres formantes

El sistema de la Figura 2.2 puede describirse usando teoría acústica y mediante técnicas numéricas usarlo para realizar simulación de generación de sonidos y su transmisión en el tracto vocal, sin embargo, por ahora es suficiente modelar la producción de un segmento de señal mediante un modelo simplificado del tracto vocal como el que se muestra en la Figura 2.3

El sistema lineal discreto variante en el tiempo de la parte derecha de la Figura 2.3 simula el moldeado en la frecuencia que realiza el tracto vocal. El generador de excitaciones de la parte izquierda simula los diferentes modos de generación de sonido del tracto vocal. Como el tracto vocal cambia relatívamente despacio, parece razonable asumir que la respuesta del sistema lineal no cambia en menos de 10ms aproximadamente.

$$H(z) = \frac{\sum_{k=0}^{M} b_b z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{2.1}$$

donde los coeficientes del filtro a_k y b_k, etiquetados como "parámetros del tracto vocal" en la Figura 2.2 cambian a un régimen de 50 a 100 veces por segundo. Algunos de los polos c_k de la función sistema residen dentro del círculo trigonométrico unitario y generan resonancias que corresponden con las frecuencias formantes. En el modelado detallado de la producción

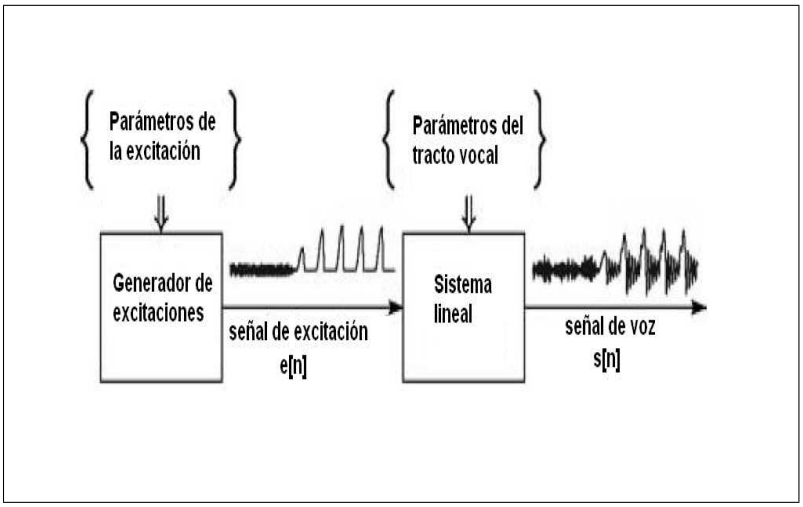

Figura 2.3: Modelo simplificado del tracto vocal

de voz, a veces es de utilidad emplear ceros d_k de la función sistema para modelar sonidos fricativos y nasales. Sin embargo, mucho modelos solo emplean polos en el modelo para simplificar el análisis requerido para estimar los parámetros del modelo. La caja etiquetada "Generador de excitaciones" en la figura 2.3 crea una excitación apropiada para el tipo de sonido producido. Para sonidos vocalizados, la excitación del sistema es una secuencia cuasi-periódica de pulsos glotales discretos que se parecen mucho a los mostrados en la Figura 2.3. La frecuencia fundamental de la excitación glotal determina el tono percibido de la voz. Los pulsos glotales individuales tienen un espectro de bajo espectro debido a varios factores. La secuencia periódica de pulsos glotales suavizados tienen un espectro cuyos componentes disminuyen en amplitud a medida que aumenta la frecuencia. A veces conviene incluir la contribución del espectro del pulso glotal en el modelo mediante un pequeño incremento en el orden del denominador respecto a lo necesario para representar las frecuencias formantes. Para voz no-vocalizada, el sistema lineal es excitado por un generador de números aleatorios que produce una señal de ruido discreto con espectro plano. En la Figura 2.3, la excitación cambia de ser no-vocalizada a vocalizada. En cualquier caso el sistema lineal impone su respuesta a la frecuencia para crear sonidos.

La señal de voz puede ser representada por los parámetros del modelo en lugar de la forma de onda muestreada. Asumiendo que las propiedades de la señal de voz (y del modelo) son constantes por intervalos de tiempo cortos, es

posible calcular, medir o estimar los parámetros del modelo analizando segmentos cortos de la señal de voz y de esa manera obtener una representación digital de la señal.

2.4. Teoría Acústica

Portnoff demostró que las ondas de sonido se propagan en un tubo de acuerdo a las siguientes ecuaciones [3]:

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial (u/A)}{\partial t} \tag{2.2}$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2} \frac{\partial (pA)}{\partial t} + \frac{\partial A}{\partial t} \tag{2.3}$$

donde:
$p = p(x,t)$ es la variación de la presión del sonido dentro del tubo en la posición x e instante t.
$u = u(x,t)$ es la variación de la velocidad del flujo en la posición x e instante t.
ρ es la densidad de aire en el tubo.
c es la velocidad del sonido.
$A = A(x,t)$ es el área de sección transversal del tubo en la posición x e instante t.

Encontrar soluciones cerradas de estas ecuaciones no es posible excepto para los casos mas simples. Se pueden obtener soluciones numéricas, pero se tiene que determinar previamente de alguna manera la presión y la velocidad del volumen de aire en el glotis y en los labios para cada instante de tiempo, es decir de las "condiciones de frontera", para ello en cuanto a los labios se debe tomar en cuenta la radiación del sonido y en cuanto al glotis se debe tomar en cuenta la naturaleza de la excitación. Adicionalmente, la función de área $A(x,t)$ debe ser conocida y como es de imaginarse, mediciones detalladas son muy difíciles de obtener, aunque hay quien lo ha intentado a partir de secuencias de imágenes de rayos X.

2.4.1. Propagación del sonido en un tubo uniforme sin pérdidas. Modelado a partir de ecuaciones diferenciales

Para acercarse al entendimiento de la naturaleza de la propagación del sonido en el tracto vocal se puede considerar primero el caso mas simple, es decir, el caso en el que el área de sección transversal $A(x,t)$ es constante tanto en x como en t, el tubo es excitado por una fuente ideal de velocidad de flujo representada por un pistón, adicionalmente podemos asumir que a lo largo del tubo no hay cambios de presión, solo de velocidad de flujo, en este caso, las ecuaciones se reducen a:

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial u}{\partial t} \tag{2.4}$$

$$-\frac{\partial u}{\partial x} = \frac{A}{\rho c^2}\frac{\partial p}{\partial t} + \frac{\partial A}{\partial t} \tag{2.5}$$

Se puede demostrar que la solución de este sistema de ecuaciones diferenciales tiene la forma:

$$u(x,t) = [u^+(t - x/c) - u^-(t + x/c)] \tag{2.6}$$

$$p(x,t) = \frac{\rho c}{A}[u^+(t - x/c) + u^-(t + x/c)] \tag{2.7}$$

Las funciones $u^+(t-x/c)$ y $u^-(t+x/c)$ se pueden interpretar como ondas viajando en la dirección positiva y negativa respectivamente.

Es conveniente hacer la analogía entre la propagación del sonido en un tubo sin pérdidas y la propagación de la corriente eléctrica en una línea de transmisión sin pérdidas, donde el voltaje $v(x,t)$ y la corriente eléctrica $i(x,t)$ satisfacen las ecuaciones:

$$-\frac{\partial v}{\partial x} = L\frac{\partial i}{\partial t} \tag{2.8}$$

$$-\frac{\partial i}{\partial x} = C\frac{\partial v}{\partial t} \tag{2.9}$$

donde L y C son la inductancia y la capacitancia por unidad de longitud respectivamente. La siguiente tabla resume la analogía:

Tabla 2.1: Analogías entre variables acústicas y eléctricas

Variable acústica	Variable Eléctrica
p presión	v voltaje
u velocidad de volumen	i corriente
ρ/A inductancia acústica	L inductancia
$A/\rho c^2$ Capacitancia acústica	C capacitancia

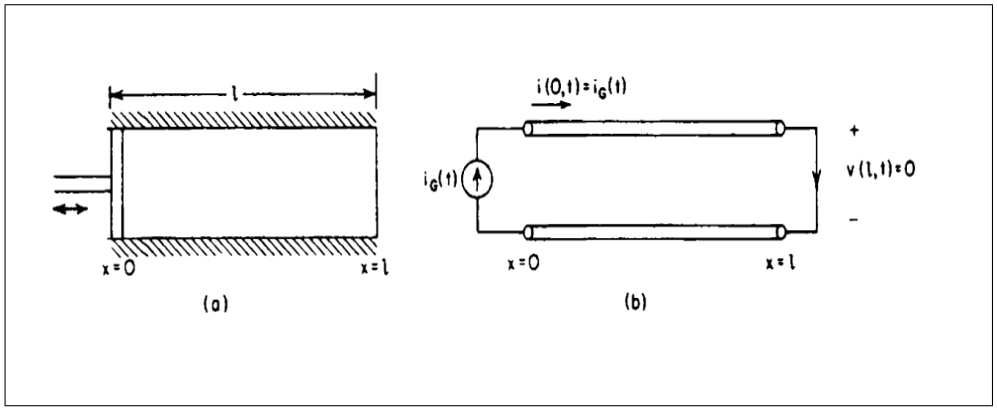

Figura 2.4: Analogía entre un tubo uniforme sin pérdida y una línea de transmisión Eléctrica ideal

Usando las analogías de la Tabla 2.1, podemos decir que el tubo de sección transversal uniforme al que un pistón inyecta aire de la Figura 2.4(a) se comporta en forma idéntica a una línea de transmisión terminada en corto-circuito ($V(l,t) = 0$) y excitada por una fuente de corriente como se muestra en la Figura 2.4(b).

2.5. Modelo del tracto vocal basado en tubos sin pérdidas concatenados

Uno de los modelos del tracto vocal mas utilizado se basa en la concatenación de de tubos uniformes sin pérdidas de diferentes longitudes y áreas de sección transversal como el modelo de cinco tubos mostrado en la Figura 2.5

Como cada segmento de tubo es uniforme y sin pérdidas, las ecuaciones

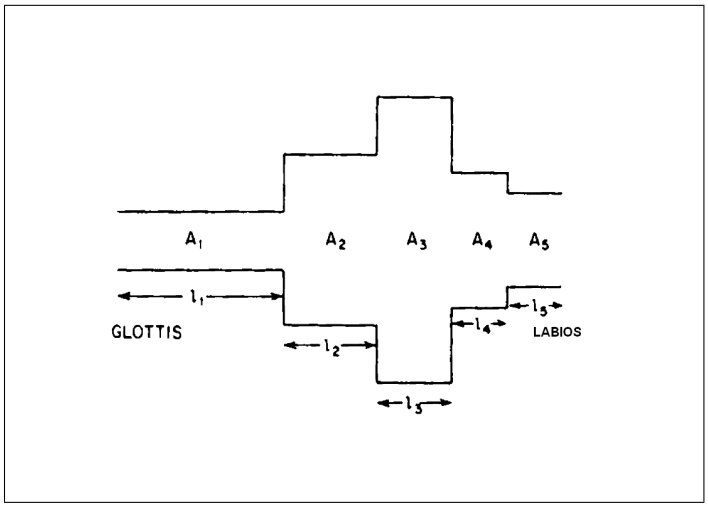

Figura 2.5: Modelo del Tracto Vocal basado en la Concatenación de 5 tubos uniformes sin pérdidas

que describen cada segmento son las mismas que (2.6) y (2.7) pero con valores apropiados para el segmento en cuestión, entonces, para el k-ésimo tubo con área de sección transversal A_k y longitud l_k la velocidad de propagación de la onda y la presión serán:

$$u_k(x,t) = [u_k^+(t - x/c) - u_k^-(t + x/c)] \tag{2.10}$$

$$p_k(x,t) = \frac{\rho c}{A_k}[u_k^+(t - x/c) + u_k^-(t + x/c)] \tag{2.11}$$

donde x es la distancia medida desde el extremo izquierdodel k-ésimo tubo ($0 \leq x \leq l_k$.

2.5.1. Determinación de los coeficientes de reflexión de la onda de sonido a partir de las condiciones de frontera

El que dos tubos adyacentes estén conectados genera las condiciones de adyacencia en la que la presión al final de un tubo es igual que la presión al inicio del siguiente tubo, es decir, algo similar podemos decir respecto a la velocidad de propagación de un volumen:

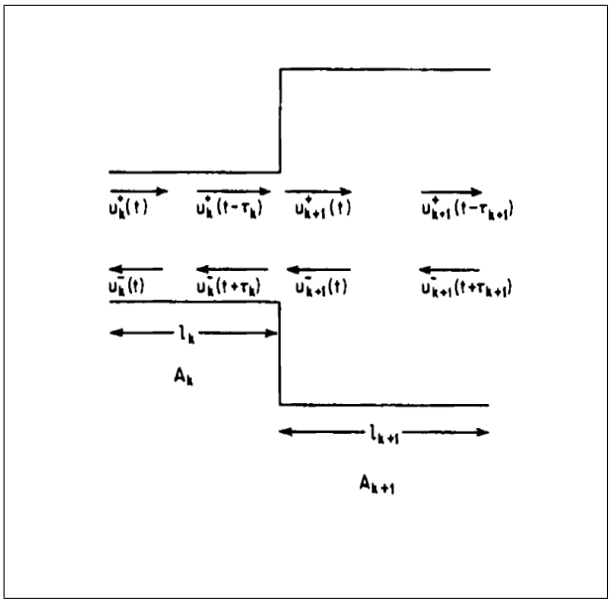

Figura 2.6: Notación utilizada para representar ondas propagándose en tubos uniformes concatenados

$$p_k(l_k, t) = p_{k+1}(0, t) \tag{2.12}$$

$$u_k(l_k, t) = u_{k+1}(0, t) \tag{2.13}$$

Sustituyendo estas en las ecuaciones (2.11) y (2.10), obtenemos:

$$\frac{A_{k+1}}{A_k}[u_k^+(t-\tau_k) + u_k^-(t+\tau_k)] = [u_{k+1}^+(t) + u_{k+1}^-(t)] \tag{2.14}$$

$$u_k^+(t-\tau_k) - u_k^-(t+\tau_k) = u_{k+1}^+(t) + u_{k+1}^-(t) \tag{2.15}$$

donde $\tau_k = l_k/c$ representa el tiempo que le toma a la onda acústica atravesar el tubo de longitud l_k.

$u_{k+1}^+(t)$ representa la onda propagándose a la derecha al inicio del segmento $k+1$ mientras que $u_k^-(t+\tau_k)$ es la onda propagándose a la izquierda al final del segmento k, vea la Figura 2.6.

Al propagarse hacia la derecha, la onda eventualmente llega al extremo al final del segmento k ($u_k^+(t-\tau_k)$) donde este se une con el segmento $k+1$, una parte de la onda continúa propagándose hacia la derecha ($u_{k+1}^+(t)$) y otra

parte es reflejada hacia la izquierda ($u_k^-(t+\tau_k)$). Despejando $u_k^-(t+\tau_k)$ de (2.15) y sustituyendo en (2.14) obtenemos:

$$u_{k+1}^+(t) = \frac{2A_{k+1}}{A_{k+1}+A_k}u_k^+(t-\tau_k) + \frac{A_{k+1}-A_k}{A_{k+1}+A_k}u_{k+1}^-(t) \tag{2.16}$$

Ahora, restando miembro a miembro (2.15) de (2.14) obtenemos:

$$u_k^-(t+\tau_k) = -\frac{A_{k+1}-A_k}{A_{k+1}+A_k}u_k^+(t-\tau_k) + \frac{2A_k}{A_{k+1}+A_k}u_{k+1}^-(t) \tag{2.17}$$

La cantidad:

$$r_k = \frac{A_{k+1}-A_k}{A_{k+1}+A_k} \tag{2.18}$$

es la fracción de $u_k^-(t)$ que es reflejada en la unión, por lo cual se le denomina "coeficiente de reflexión de la k-ésima unión". En vista de que todas las áreas de sección transversal son necesariamente positivas se puede demostrar que $-1 < r_k < 1$. Las ecuaciones de propagación de las ondas acústicas al inicio del k+1-ésimo tubo y en el k-ésimo tubo las podemos expresar entonces en términos del coeficiente reflejante de la k-ésima unión el cual a la vez depende de la relación que hay entre las áreas de sección transversal de los tubos que se conectan en dicha unión.

$$u_{k+1}^+(t) = (1+r_k)u_k^+(t-\tau_k) + r_k u_{k+1}^-(t) \tag{2.19}$$

$$u_k^-(t+\tau_k) = -r_k u_k^+(t-\tau_k) + (1-r_k)u_{k+1}^-(t) \tag{2.20}$$

2.5.2. Síntesis de voz usando coeficientes reflejantes

Estas ecuaciones fueron utilizadas por Kelly and Lochbaum [3]. Para comprenderlo conviene representar estas ecuaciones gráficamente como en la Figura 2.7. En esta figura, se utilizan convenciones para grafos de flujo de señales para representar las multiplicaciones y las sumas. La Figura 2.7 muestra las relaciones entre la salida de presión y flujo de volumen del segundo tubo y la presión y flujo de volumen a la entrada del primer tubo. Un modelo basado en cinco tubos concatenados tendría cinco cajas de retardo hacia adelante y hacia atras y cuatro uniones, cada unión está caracterizada

por un coeficiente de reflexión. Para completar el modelo de propagación de una onda de sonido en dicho sistema es necesario considerar las condiciones de frontera en los labios y en el glotis.

2.6. Pulso glotal

Los sonidos de la voz pueden clasificarse como vocalizados o no-vocalizados, se requiere entonces de una fuente que pueda producir ya sea una forma de onda cuasi-periódica o bien ruido. En el caso de sonidos vocalizados, la forma de onda de la excitación debería parecerse a la de la Figura 2.8.

Una manera conveniente de representar la generación de la onda glotal se muestra en la Figura 2.9. El generador de tren de impulsos produce una secuencia de impulsos unitarios que están espaciados de acuerdo al periodo del tono, esta señal excita un sistema lineal cuya respuesta al impulso $g(n)$ tiene la forma de onda glotal deseada. Un control de ganancia A_v controla la intensidad de la excitación.

La elección de la forma de $g(n)$ no es crítica con tal de que su transformada de Fourier tenga las propiedades correctas. En un estudio hecho por Rosenberg del efecto que la forma del pulso glotal tenía en la calidad de la voz, encontró que la forma de onda del pulso glotal podía ser reemplazada por un la forma de onda del pulso sintético de la forma:

$$\begin{aligned} g(n) &= \frac{1}{2}[1 - cos(\pi n/N_1)] \quad 0 \leq n \leq N_1 && (2.21)\\ &= cos(\pi(n - N_1)/2N_2) \quad N_1 \leq n \leq N_1 + N_2 && (2.22)\\ &= 0 \quad en\ otro\ lugar && (2.23) \end{aligned}$$

El efecto del modelo del pulso glotal en el dominio de la frecuencia es introducir un filtrado pasabajas, el modelo propuesto es un modelo de puros ceros, a menudo se prefiere un modelo de puros polos, se ha conseguido modelos exitosos de dos polos para $G(z)$ [3].

2.7. Modelo completo del tracto vocal

Poniendo todos los ingredientes juntos, obtenemos el modelo de la Figura 2.10. Alternando entre los generadores de excitación vocalizada y no vocalizada cambiamos los modos de excitación. El tracto vocal puede ser modelado

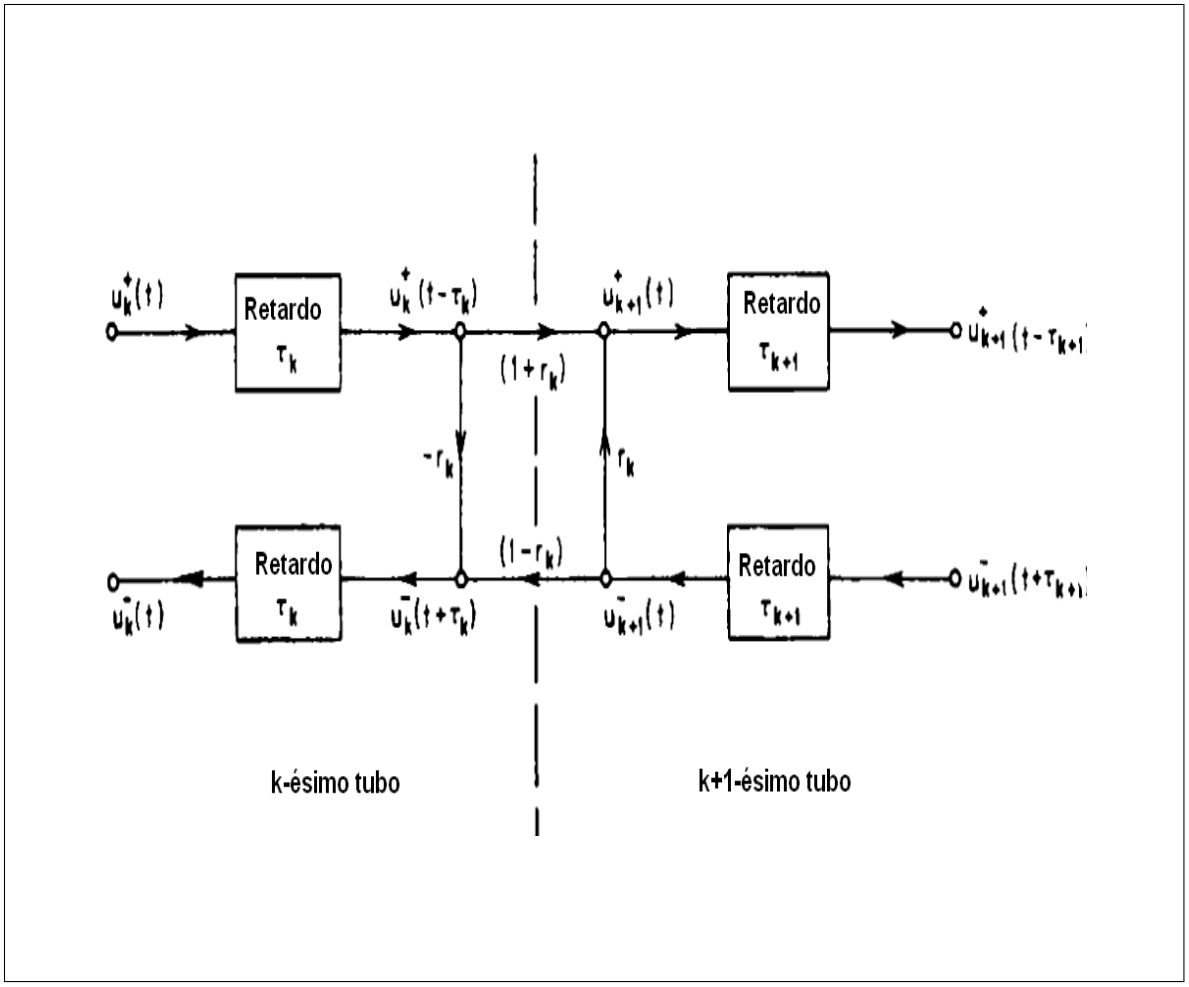

Figura 2.7: Representación gráfica de las ecuaciones de transporte de las ondas acústicas en tubos sin pérdidas concatenados

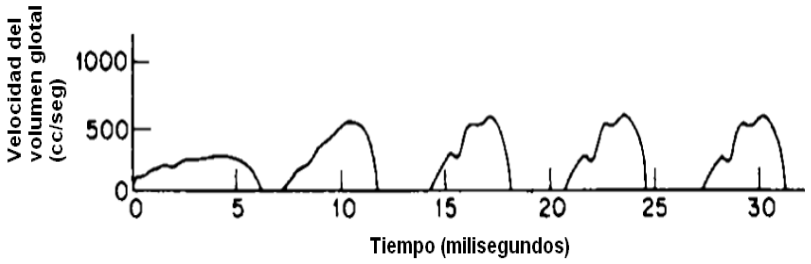

Figura 2.8: Tren de pulsos glotales

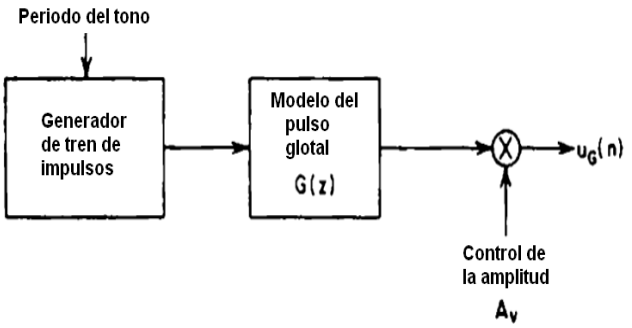

Figura 2.9: Modelo de la Excitacion

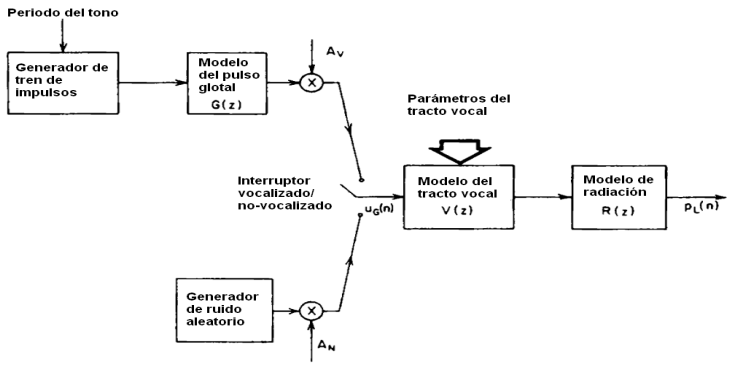

Figura 2.10: Modelo Completo del Tracto Vocal

mediante una amplia variedad de opciones. En algunos casos conviene combinar los modelos de radiación y de pulso glotal en un solo sistema. De hecho como veremos mas adelante, en el caso del modelo basado en un predictor lineal se combinan el pulso glotal, la radiación y el tracto vocal en una sola función de transferencia de puros polos.

El modelo completo del tracto vocal tiene sus deficiencias pero afortunadamente, ninguna de ellas limita sus aplicaciones. La variación temporal de sus parámetros en sonidos vocalizados es bastante lenta por lo que el modelo funciona muy bien. Para otros sonidos como los plosivos, el modelo no es tan bueno pero es aún adecuado. El modelo asume que los parámetros son constantes para intervalos de tiempo entre 10 y 20 ms. Otra limitante del modelo es la ausencia de ceros que en teoría se requieren para sonidos nasales, sin embargo, el modelo se puede ampliar para incluir ceros. Finalmente, la excitación es inadecuada para fricativos vocalizados ya que estos no se pueden clasificar simplemente como vocalizados ni como no-vocalizados, existen modelos mas sofisticados que se pueden emplear para salvar esta limitación.

Capítulo 3

Procesamiento de la señal de voz en el dominio del tiempo

En el dominio del tiempo se pueden extraer ciertas características como la energía de tiempo corto, el régimen de cruces por cero de tiempo corto, la autocorrelación de tiempo corto, la entropía de tiempo corto y otras que nos permiten detectar si una señal de audio contiene voz o solo ruido de fondo (energía y régimen de cruces por cero), si la señal contiene voz, podemos determinar si se trata de sonidos vocalizados o no-vocalizados (autocorrelación) y en caso de tratarse de sonidos vocalizados determinar el tono (autocorrelación modificada), incluso se puede realizar reconocimiento de voz con características extraídas en el dominio del tiempo (usando cruces por cero de orden superior). La extracción de características en el dominio del tiempo tiene la enorme ventaja de poderse realizar muy rápidamente y por ende llevarse a cabo con dispositivos de bajo costo.

3.1. Energía de tiempo corto

La energía E de una señal x se define como:

$$E = \sum_{m=-\infty}^{\infty} x^2(m) \tag{3.1}$$

Sin embargo, pocas veces nos resulta de interés conocer la energía contenida en toda la señal de audio, en cambio, la manera en que la energía contenida en la señal de audio aumenta o disminuye es de mucha utilidad.

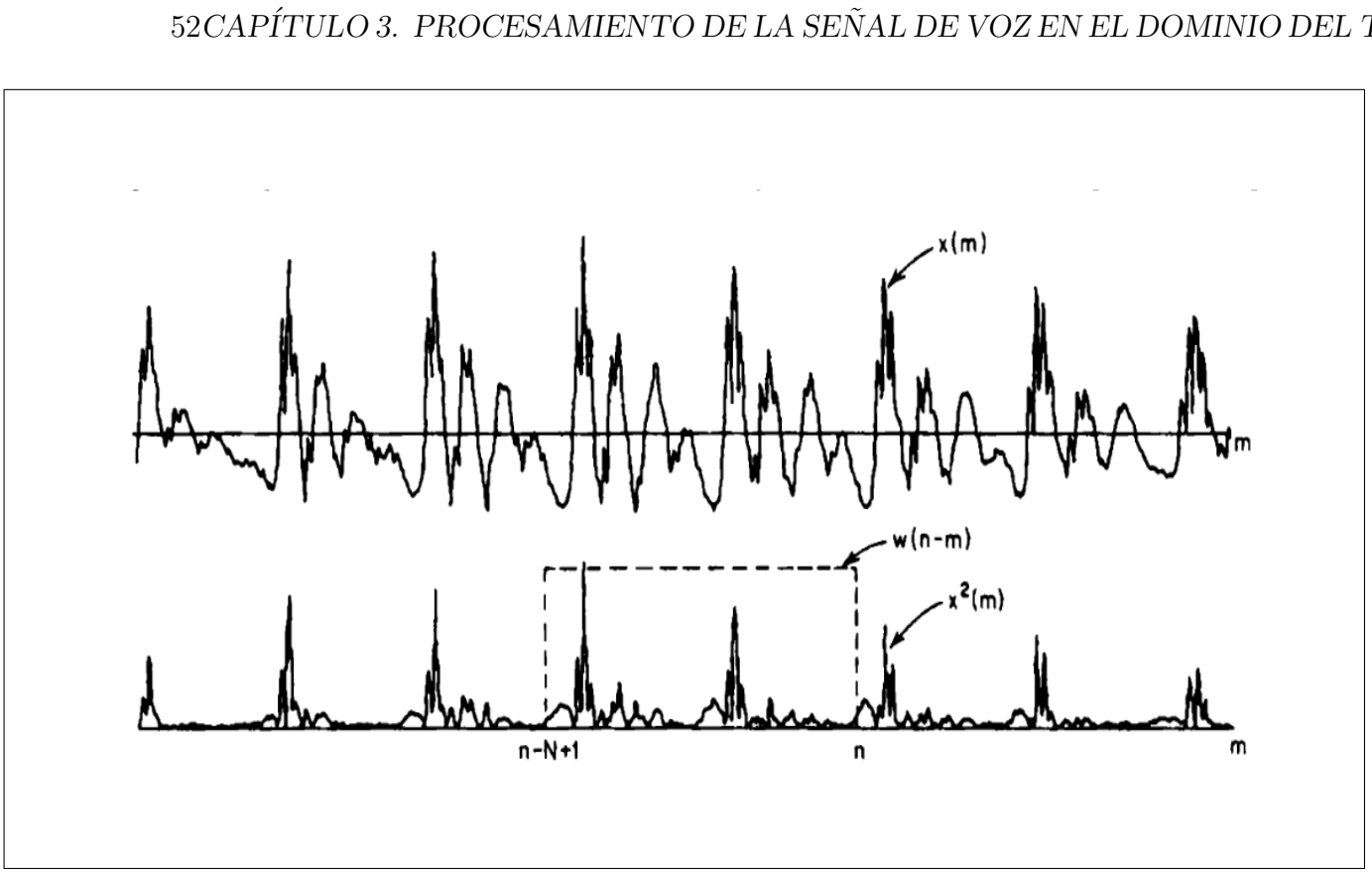

Figura 3.1: Calculando la Energía de tiempo corto

Lo mas parecido a la energía instantánea es la denominada *energía de tiempo corto* E_n que es la energía contenida en un segmento de la señal que comienza en el tiempo discreto n y se determina mediante:

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) \tag{3.2}$$

donde N es el tamaño del segmento corto de audio denominado *marco*.

Podemos ver a la señal de audio a través de una ventana de tamaño N que se desliza sobre la señal y que solo nos permite ver un segmento corto de la señal de audio a la vez. La ventana la podemos definir mediante:

$$w(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \textit{en otro lado} \end{cases} \tag{3.3}$$

En la Fig. 3.1, la señal $x(m)$ se eleva al cuadrado (observe en la parte inferior como ya no hay valores negativos) y luego se ve a través de una ventana (en realidad marco) de tamaño N, de manera que los únicas muestras que se van a sumar para determinar la energía de tiempo corto son las que se ubican dentro de la ventana, esta energía la denotamos E_n puesto que corresponde a una ventana ubicada en la posición n. una vez definida la ventana w, la energía de tiempo corto la podemos determinar mediante:

$$E_n = \sum_{m=-\infty}^{m=\infty} [x(m)w(n-m)]^2 \tag{3.4}$$

o bien mediante:

$$E_n = \sum_{m=-\infty}^{m=\infty} [x(m)^2 h(n-m)] \tag{3.5}$$

donde $h(n) = w^2(n)$

También podemos determinar la *Magnitud de tiempo corto* M_n, esto lo hacemos mediante:

$$M_n = \sum_{m=-\infty}^{m=\infty} [x(m)w(n-m)] \tag{3.6}$$

3.2. Régimen de cruces por cero de tiempo corto

Un *cruce por cero* ocurre cuando dos muestras consecutivas de la señal tienen signos distintos. El régimen al cual ocurren los cruces por cero es una medida del contenido de frecuencia en la señal, por ejemplo, una senoide de frecuencia F_0 muestreada a una frecuencia F_s será digitalizada de manera que habrá F_s/F_0 muestras por periodo y como hay dos cruces por cero en cada periodo tendremos un régimen de cruces por cero $Z = 2F_0/F_s$. Una definición adecuada del régimen de cruces por cero de tiempo corto es:

$$Z_n = \sum_{m=-\infty}^{m=\infty} [|signo[x(m)] - signo[x(m-1)]|w(n-m)] \tag{3.7}$$

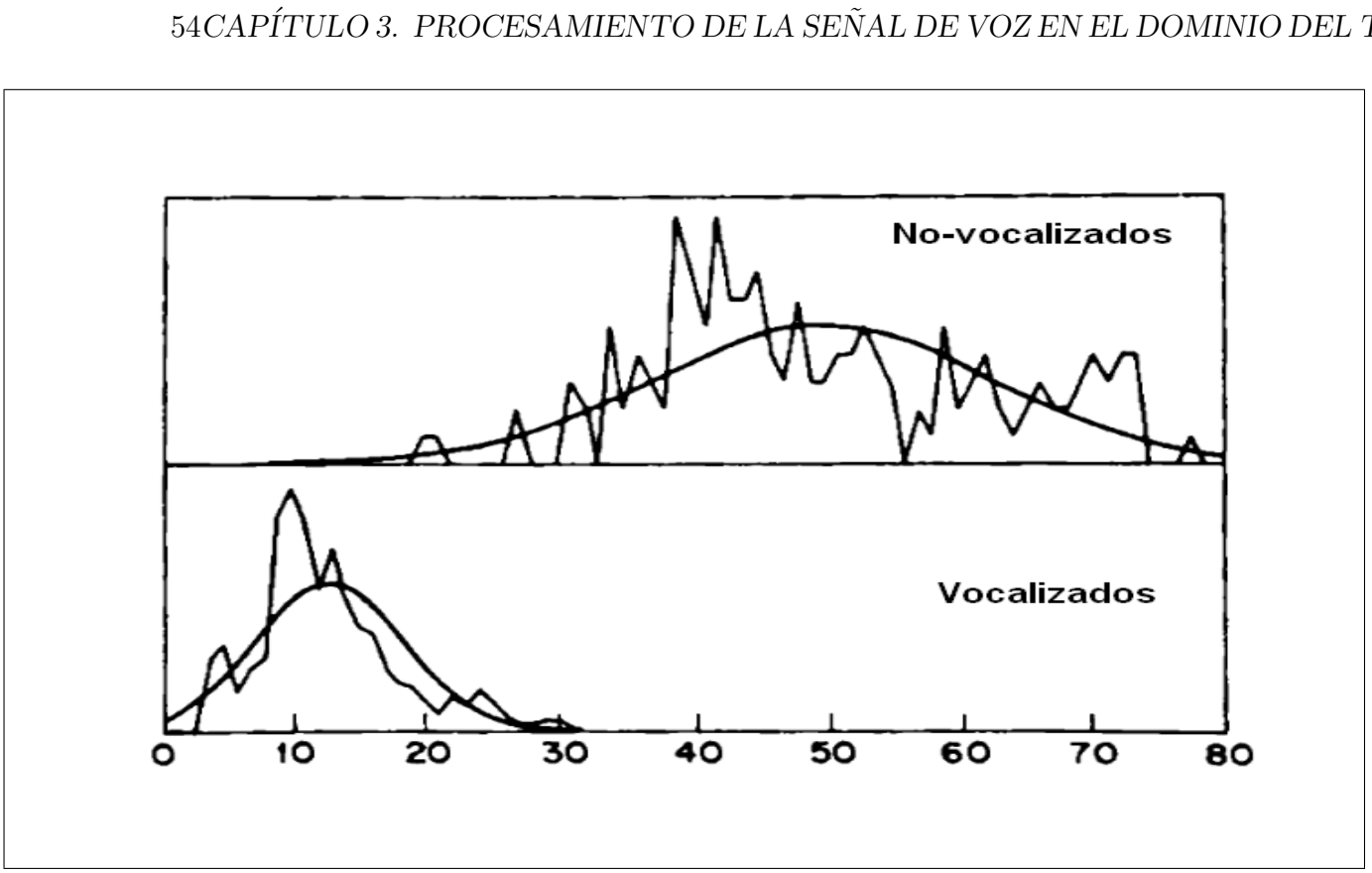

Figura 3.2: Distribución del régimen de cruces por cero en sonidos vocalizados y en sonidos no-vocalizados

donde:

$$signo[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases} \tag{3.8}$$

además:

$$w(n) = \begin{cases} \frac{1}{2N} & 0 \leq n \leq N-1 \\ 0 & en\ otro\ lado \end{cases} \tag{3.9}$$

Un uso que podemos dar al régimen de cruces por cero de tiempo corto es el de discriminar sonidos vocalizados de sonidos no-vocalizados ya que como podemos ver en la Figura 3.2 es que en general, los sonidos vocalizados tiene un régimen de cruces por cero bajo comparado con el régimen de cruces por cero de los sonidos no-vocalizados.

3.3. Estimación del tono

La estimación del tono se puede hacer mediante un enfoque de procesamiento paralelo que fácilmente puede implementarse en hardware. La señal de voz es procesada conviertiendola en un tren de impulsos que retiene la periodicidad de la señal original descartando características irrelevantes para el fin de la determinación del tono, varios estimadores del tono se combinan lógicamente para obtener una medición confiable, el procedimiento es el siguiente:

1. Pasar la señal por un filtro pasabajas con una frecuencia de corte de 900Hz, esto asegura una forma de onda suave, se puede optar por un pasabanda 100Hz-900Hz

2. Localizar picos y valles (máximos y mínimos locales) obteniendo tanto sus posiciones como sus amplitudes

3. Producir seis trenes de impulsos de acuerdo a:

   impulsos de amplitud igual a la de los picos ubicados en donde se ubican el picos.

   impulsos de amplitud igual a la diferencia entre la amplitud de cada pico y la amplitud del valle que le precede ubicados en donde se ubican los picos

   impulsos de amplitud igual a la diferencia entre la amplitud de cada pico y la amplitud del pico que le precede ubicados en donde se ubican los picos (si la diferencia es negativa la amplitud del impulso es cero)

   impulsos de amplitud igual al negativo de la amplitud de los valles ubicados en donde se ubican los valles

   impulsos de amplitud igual al negativo de la amplitud de cada valle mas la amplitud del pico que le precede ubicados donde se ubican los valles

   impulsos de amplitud igual al negativo de la amplitud de cada valle mas la amplitud del valle que le precede ubicado en cada valle (si la diferencia es negativa, la amplitud del impulso será cero)

4. Para cada tren de pulsos, se implementa un estimador simple de periodo. Cuando un impulso rebasa mayor de cierto umbral, cada vez

que uno de estos impulsos es detectado, se genera un pulso que dura al menos un tiempo τ durante el cual se ignora cualquier otro impulso, una vez terminado ese tiempo, el pulso decae exponencialmente y cualquier impulso que lo supere reinicia el pulso con la amplitud del nuevo impulso. Este procedimiento produce un tren de pulsos cuasiperiódico, la longitud de los pulsos es la estimación del tono.

5. De las seis estimaciones de tono, la moda es declarada como la estimación del tono final

6. Para sonidos no-vocalizados no hay ninguna consistencia entre los seis estimadores y de esa manera se puede identificar la naturaleza de sonido (vocalizado/no-vocalizado)

3.4. Entropía de la señal de voz

El contenido de información de un mensaje es desde el punto de vista de Claude Shannon proporcional al nivel de sorpresa obtenido en el lector, si es fácil adivinar el contenido no habrá mucha información ahí. Shannon ligó de manera irreversible el concepto de entropía o desorden de las moléculas en los gases con el de contenido de información de una señal al utilizar la misma fórmula para la "información propia"que la que utilizaba Boltzman para medir la entropía. Sean $v_1, v_2, ..., v_n$ los posible valores que pueden tomar las muestras de una señal de audio. Si por ejemplo, el tamaño de las muestras fuera de 8 bits, entonces los posibles valores serían los enteros dentro del intervalo. Digamos que v_i tiene la probabilidad p_i de ocurrir y la secuencia $p_1, p_2, ..., p_n$ es la función de Distribución de Probabilidades (PDF por sus siglas en inglés) sujeta a:

$$\sum_{i=1}^{n} p_i = 1 \tag{3.10}$$

El contenido de información I en un valor v_i también llamada "información propia"depende exclusivamente de su probabilidad $p_i = P(v_i)$ de ocurrir, para recordar esto la denotamos $I(p_i)$. Mientras menos probable sea que un valor se presente, mayor será la información que traería consigo, entonces, la información propia debe de ser una función monotónicamente decreciente de la probabilidad, además $I(p_i)$ debe de ser calculada de manera que si v_i

depende de dos o más eventos independientes con probabilidades $p_{i1}, p_{i2}, ...$, entonces la contribución al contenido de información de cada evento debe ser tal que la suma total coincida con $I(p_i)$ para poderse manejar como información, así si $p = p_{i1}p_{i2}, ...$, entonces, $I(p_i) = I(p_{i1}) + I(p_{i2}), ...$ Estas restricciones entre otras llevaron a Shannon a la conclusión de que la única función que podía cumplirlas era la función logarítmica, y que la información propia debe de calcularse mediante (3.11) [4], la base del logaritmo no es importante, Shanon utilizó base 2 por que lo que le interesaba a el era determinar la codificación óptima con ceros y unos de mensajes a transmitirse por una red.

$$I(p_i) = ln\Big(\frac{1}{p_i}\Big) = -ln(p_i) \tag{3.11}$$

La Entropía H es igual a la esperanza matemática del contenido de información en una secuencia, es decir, es el promedio de los contenidos de información de todos los valores posibles ponderado por sus probabilidades de ocurrir, tal y como lo indica la ecuación (3.12), al equivalente continuo se le conoce como "Entropía diferencial". Como al entropía de una señal es una medida de lo impredecible que esta es, resulta mínima para una señal constante de valor k, su PDF es un impulso unitario ubicado en k ,es decir, $p_i = \delta(k)$, y su entropía es cero como se muestra en (3.13), en el otro extremo la entropía es máxima para la distribución uniforme cuya PDF es $p_i = 1/n$ para n valores posibles, la entropía máxima es $log(n)$ como en (3.14).

$$H(x) = E[I(p)] = \sum_{i=1}^{n} p_i I(p) = -\sum_{i=1}^{n} p_i ln(p_i) \tag{3.12}$$

$$H_{min} = -\sum_{i} \delta(k) ln[\delta(k)] = -ln(1) = 0 \tag{3.13}$$

$$H_{max} = -\sum_{i} \frac{1}{n} ln(\frac{1}{n}) = -ln(\frac{1}{n}) = ln(n) \tag{3.14}$$

Si por ejemplo el tamaño de las muestras fuera de 16 bits, la entropía máxima sería de 11.09 ($ln(2^{16})$), sin embargo, para una señal real de audio ese nivel de entropía es casi imposible puesto que se necesitaría que cada posible valor apareciera el mismo número de veces, por ejemplo para un segmento de 2.972 segundos de audio muestreado a 44,100 muestras por segundo cada posible valor debe aparecer exactamente dos veces.

Para calcular la entropía de una señal se debe hacer una estimación de p_i para todo v_i , para estimar la función de densidad de probabilidades se puede construir el histograma de la señal y luego usar las ecuaciones (3.15) y (3.16), sin embargo, en la construcción del histograma se deben involucrar suficientes muestras para lograr una estimación confiable por ejemplo si se calcula la entropía de una señal de audio con un tamaño de muestra de 8 bits y un régimen de muestreo de 44,100 muestras por segundo para un segmento de audio de dos segundos se contaría con 88,200 valores para repartirlos en una tabla de solo 256 entradas y seguramente se estaría haciendo una excelente estimación de la PDF, además el histograma se puede mantener actualizado de manera que cada vez que se lea una muestra nueva de la señal de audio simplemente se incremente en uno la frecuencia correspondiente a ese valor y se decremente en uno el valor de la muestra que sale de la ventana de análisis. Si no se puede garantizar la suficiencia de muestras, por ejemplo para una señal de audio de solo 8,000 muestras por segundo y 16 bits por muestra se contaría con solo 240 valores para repartirlos en una tabla de 65,535 entradas si la ventana de análisis es de solo 30 milisegundos lo cual es típico cuando se procesa la señal de voz. En los métodos paramétricos, primero de debe asumir el tipo de distribución al que se apega la señal, una vez elegido el tipo de distribución se deben estimar sus parámetros. Si por ejemplo, se asume que la señal de audio sigue una distribución gaussiana, se puede utilizar la ecuación (3.17) que es la entropía de una variable aleatoria con distribución $N(0, R)$. En los métodos no paramétricos no es necesario asumir nada respecto al tipo de distribución ni estimar parámetros la función de densidad de probabilidades es moldeada por los datos y después suavizada por un kernel en un proceso iterativo hasta que se estabiliza, el más popular de estos métodos es el de la ventana de Parzen [5], sin embargo, los métodos no paramétricos son costosos computacionalmente y por ende no muy usados en aplicaciones de tiempo real.

$$p_i = \frac{f_i}{N} \tag{3.15}$$

donde f_i es el número de veces que el valor v_i ocurre en la señal x de acuerdo a (3.16).

$$f_i = \sum_{j=1}^{N} \varphi(x_j, v_i) \tag{3.16}$$

donde $\varphi(x, y) = 1$ si $x = y$ y $\varphi(x, y) = 0$ en caso contrario

$$H = \frac{n}{2} ln(2\pi) + \frac{1}{2} ln(|R|) \tag{3.17}$$

Donde R es la matriz de covarianzas de grado n .

3.5. Segmentación de palabras aisladas

Sin embargo, el uso mas importante del régimen de cruces por cero está en la segmentación de palabras aisladas y mas específicamente en la discriminación entre silencio y la voz. En la Figura 3.3 se observa como se logra la detección del inicio de la elocución de la palabra *six*, monitorando el régimen de cruces por cero de tiempo corto, el cual se eleva respecto al del ruido del ambiente debido a la presencia del fricativo *s*.

En vista de que no todas las palabras comienzan con un fricativo, conviene combinar el régimen de cruces por cero de tiempo corto con la energía de tiempo corto para implementar un discriminador silencio/voz, de esta manera, si se mantiene por arriba de un cierto valor umbral el régimen de cruces por cero de tiempo corto o por encima de otro valor umbral la energía de tiempo corto, podemos considerar que la señal de audio tiene contenido de voz y solo si ambos están por debajo de sus respectivos valores umbrales podemos considerar que la señal tiene solo silencio o ruido ambiental (de poca energía y de baja frecuencia).

3.6. Autocorrelación de tiempo corto

La autocorrelación de una señal x es una función definida como:

$$\phi(k) = \sum_{m=\infty}^{\infty} x(m)x(m-k) \tag{3.18}$$

La autocorrelación de una señal es una forma de obtener ciertas propiedades de la señal, por ejemplo, si la señal es periódica con periodo P (P muestras), es fácil demostrar que:

$$\phi(k) = \phi(k + P) \tag{3.19}$$

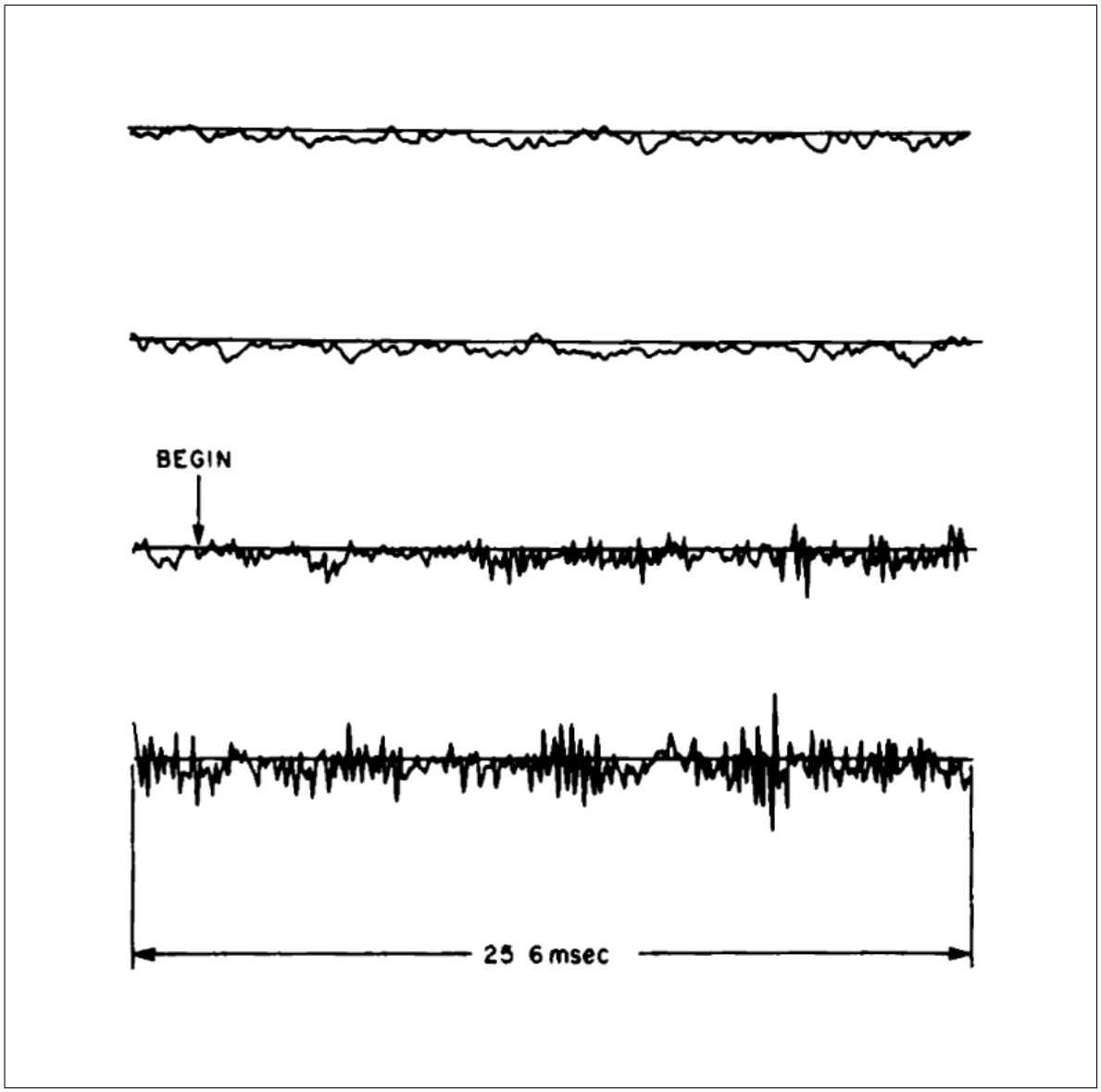

Figura 3.3: Detección del inicio de la elocución de la palabra *six* mediante régimen de cruces por cero

Por tanto, la autocorrelación de una señal periódica es a su vez periódica y con el mismo periodo. Otras propiedades importantes de la función de autocorrelación son:

1. Es una función par, es decir, $\phi(k) = \phi(-k)$
2. Tiene su máximo global en $k = 0$, es decir, $|\phi(k)| < \phi(0)$ para todo k
3. La energía de la señal es igual a $\phi(0)$

Si tomamos en cuenta las propiedades (1) y (2), podemos concluir que la función de autocorrelación tiene máximos locales en $0, \pm P, \pm 2P, ...$, por lo tanto el periodo de la señal puede ser determinado encontrando la ubicación del primer máximo local de la función de autocorrelación. Esta propiedad hace a la función de autocorrelación muy atractiva como estimador de la periodicidad de señales cortas incluida la señal de voz. La autocorrelación retiene tanta información de la señal de voz que podemos usar solo los primeros 10-15 valores de la autocorrelación de tiempo corto para a partir de ellos, podemos determinar coeficientes LPC utilizados para comprimir y descomprimir la señal. La autocorrelación de tiempo corto de la señal x se puede definir de manera similar a como definimos la energía de tiempo corto o el régimen de cruces por cero de tiempo corto:

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)x(m+k)w(n-k-m) \tag{3.20}$$

Esta ecuación se puede entender como tomar un segmento de la señal (mediante la multiplicación por la ventana) y luego determinar la autocorrelación de dicho segmento de señal. Considerando ventanas de duración N, la autocorrelación de tiempo corto se puede calcular de la siguiente forma:

$$R_n(k) = \sum_{m=0}^{N-1-k} [x(n+m)w'(m)][x(n+m+k)w'(k+m)] \tag{3.21}$$

donde $w'(n) = w(-n)$.

El cálculo de la autocorrelación tiene complejidad cuadrática pero hay métodos como el de Blankenship que aprovechan sus propiedades para ahorrar cálculos en base a una formulación recursiva de la autocorrelación [3]

En la Figura 3.4 vemos tres ejemplos de funciones de autocorrelación determinadas para una señal de voz muestreada a 10KHz usando $N = 401$. En los primeros dos casos (a) y (b) se trata de segmentos vocalizados y en el tercero (c) de un segmento no-vocalizado, en el primer segmento los picos ocurren aproximadamente en múltiplos de 72 indicando un periodo de 7.2 ms, es decir una frecuencia fundamental de 140 Hz, en el segundo caso (b) los picos ocurren en múltiplos de 58 muestras indicando un periodo de 5.8 ms. En el último caso (c) no hay picos periódicos indicando ausencia de periodicidad en la señal.

El tamaño de la ventana N debe ser lo suficientemente pequeño como para reflejar los cambios rápidos de la señal pero lo suficientemente grande como para lograr estimar la periodicidad de la señal. Como regla general, la ventana debe contener al menos dos periodos de la forma de onda.

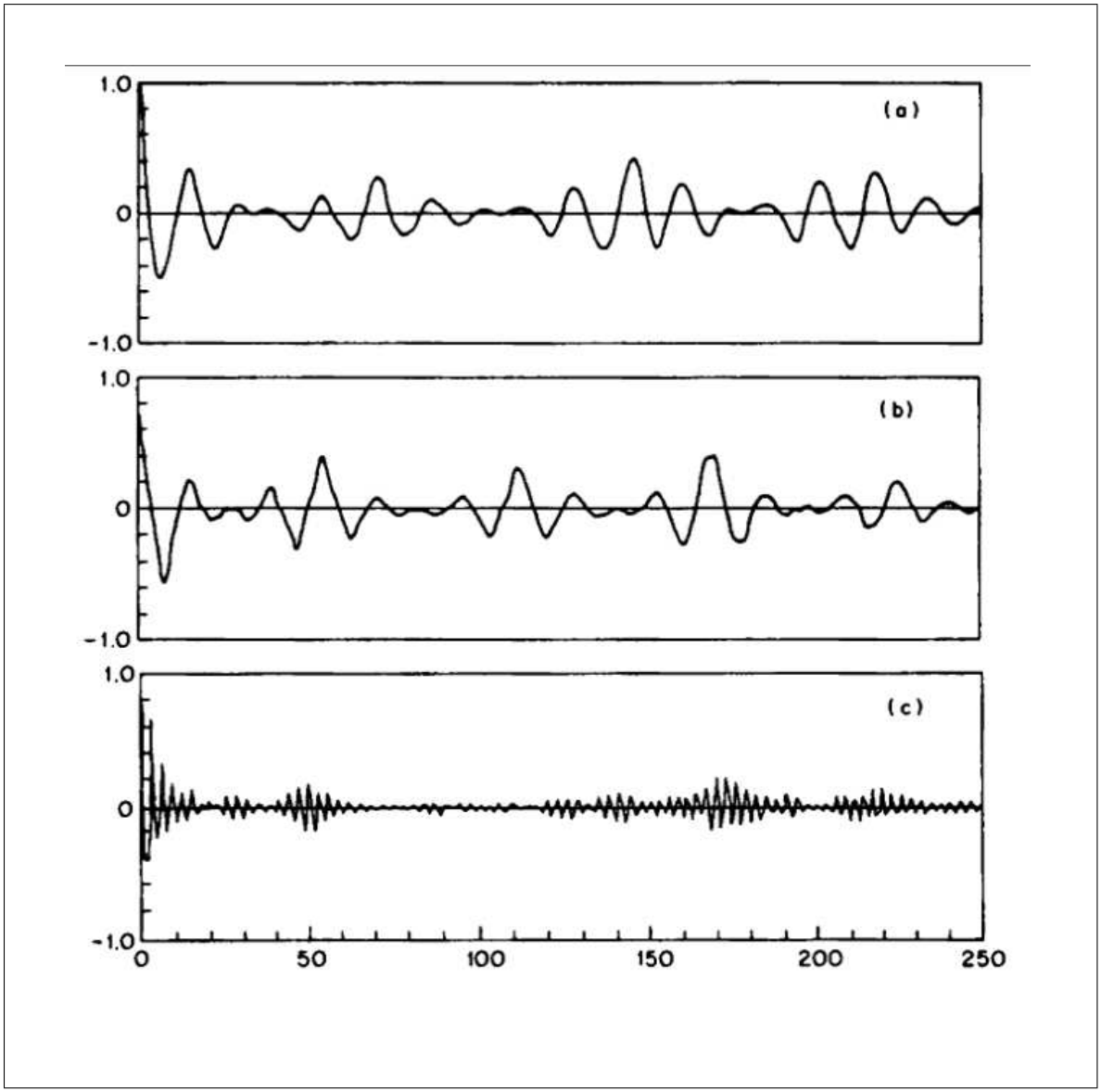

Figura 3.4: Autocorrelación de tempo corto para dos segmentos vocalizados (a) y (b) y uno no-vocalizado (c)

3.6.1. Algoritmo de discriminación silencio/voz vocalizado/no-vocalizado y estimación del tono mediante autocorrelación

En cierto sentido, la autocorrelación retiene demasiada información acerca de la señal de voz, de hecho, los primeros 10 valores de la autocorrelación son suficientes para estimar de manera precisa la función de transferencia del tracto vocal. La función de autocorrelación tiene muchos picos pero la mayoría de estos picos se atribuyen a oscilaciones del tracto vocal que son responsables de darle forma a cada periodo de la señal de voz. El procedimiento sencillo de buscar la ubicación del pico mas grande de la autocorrelación para estimar el tono no es tan confiable debido a que en ocasiones, el cambio rápido de los formantes puede crear un pico en la autocorrelación mas grande que aquel debido al tono. Debido a estas razones, conviene enfatizar la periodicidad de la señal suprimiendo al mismo tiempo de la función de autocorrelación algunas de sus características que pueden resultar distractores del problema de determinación del tono. A las técnicas que llevan a cabo esta función se les denomina *aplanadores del espectro*, uno de ellos consiste en aplicar a la señal de voz la función de *recorte del centro* tal y como se define en la Figura 3.5.

En la Figura 3.6 podemos ver el efecto que la aplicación de la función de *recorte del centro* tendría sobre una hipotética señal de voz, para muestras con valor mayor al nivel de recorte, la salida del recortador es en efecto igual a la entrada menos el nivel de recorte mientras que para muestras con valor inferior al nivel de recorte, la salida es cero.

En la Figura 3.7(a) vemos un sonido vocalizado y su función de autocorrelación y en la Figura 3.7(b) vemos al mismo sonido ya recortado y su respectiva función de autocorrelación de donde la estimación del tono se vuelve mas fácil de obtener y mas confiable.

Combinando el recorte de centro, la autocorrelación de tiempo corto y la energía de tiempo corto se puede implementar un algoritmo como el mostrado en la Figura 3.8 para determinar primero si en la señal de audio hay voz o silencio, en caso de que exista voz si se trata de sonido vocalizado o no-vocalizado. Finalmente, en caso de tratarse de sonido vocalizado determinar el tono.

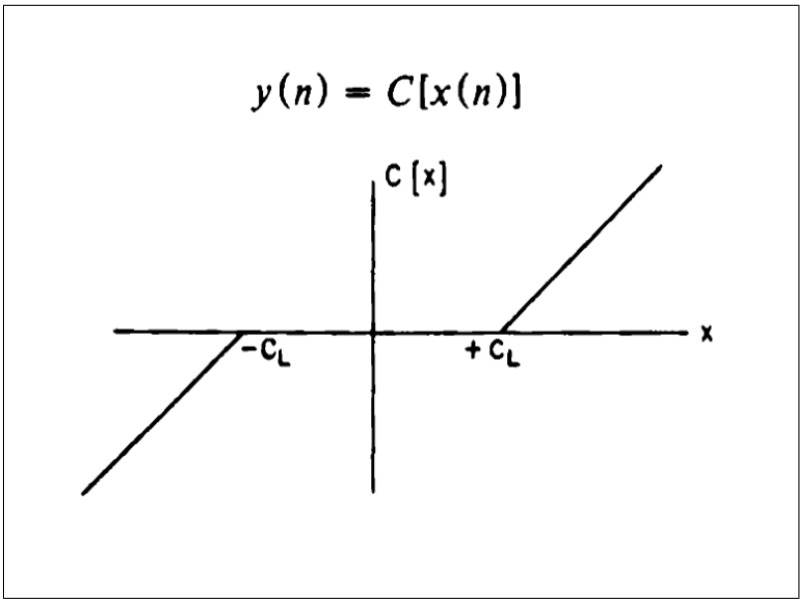

Figura 3.5: Función de recorte del centro

3.6.2. Autocorrelación Modificada de tiempo corto (Correlación cruzada entre marcos consecutivos)

Debido a la longitud finita de la ventana utilizada en el cálculo de la autocorrelación de tiempo corto, hay cada vez menos datos involucrados en el cálculo de $R_n(k)$ a medida que k aumenta, esto conduce a a una reducción en la amplitud de los picos a medida que k aumenta, este efecto se puede apreciar en la Figura 3.9.

La autocorrelación de una onda estrictamente periódica tendría picos equiespaciados de exáctamente la misma amplitud, sin embargo eso no ocurriría para la autocorrelación de tiempo corto debido al efecto que se acaba de explicar y que se muestra en la Figura 3.9 donde para la misma onda se calcula la autocorrelacion de tempo corto para diferentes tamaños de ventana, observe que para la Figura 3.9(c) el pico ubicado en $k = 72$ no es el pico mas prominente debido a que para un tamaño de ventana de 125 ya son muy pocas las muestras involucradas en la sumatoria cuando se está determinando $R_n(75)$. Para evitar este problema se puede utilizar la función de autocorrelación modificada la cual se define como:

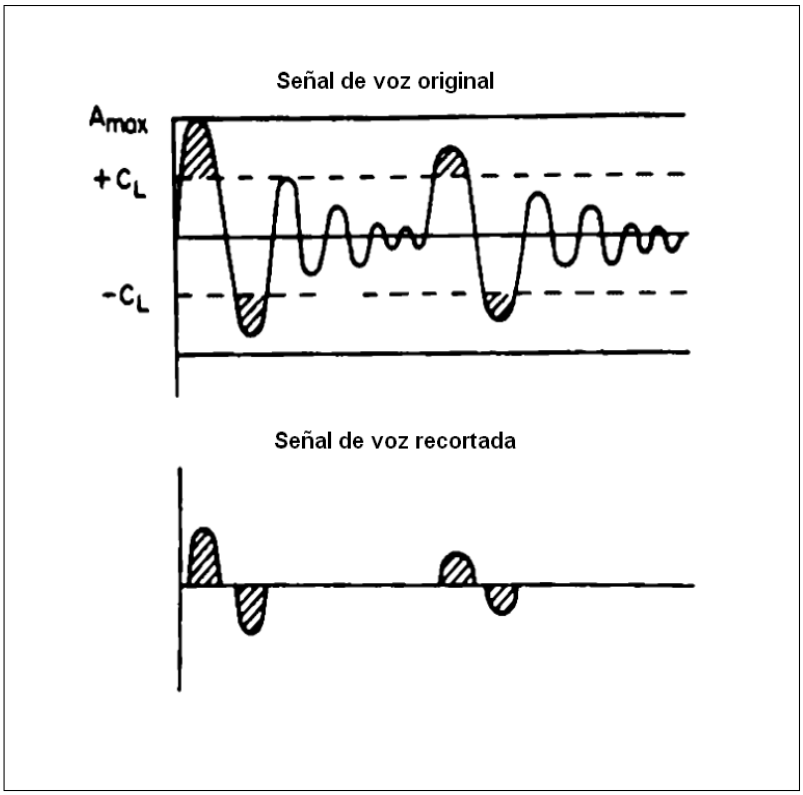

Figura 3.6: Efecto de aplicación de la función de recorte del centro sobre una señal de voz hipotética

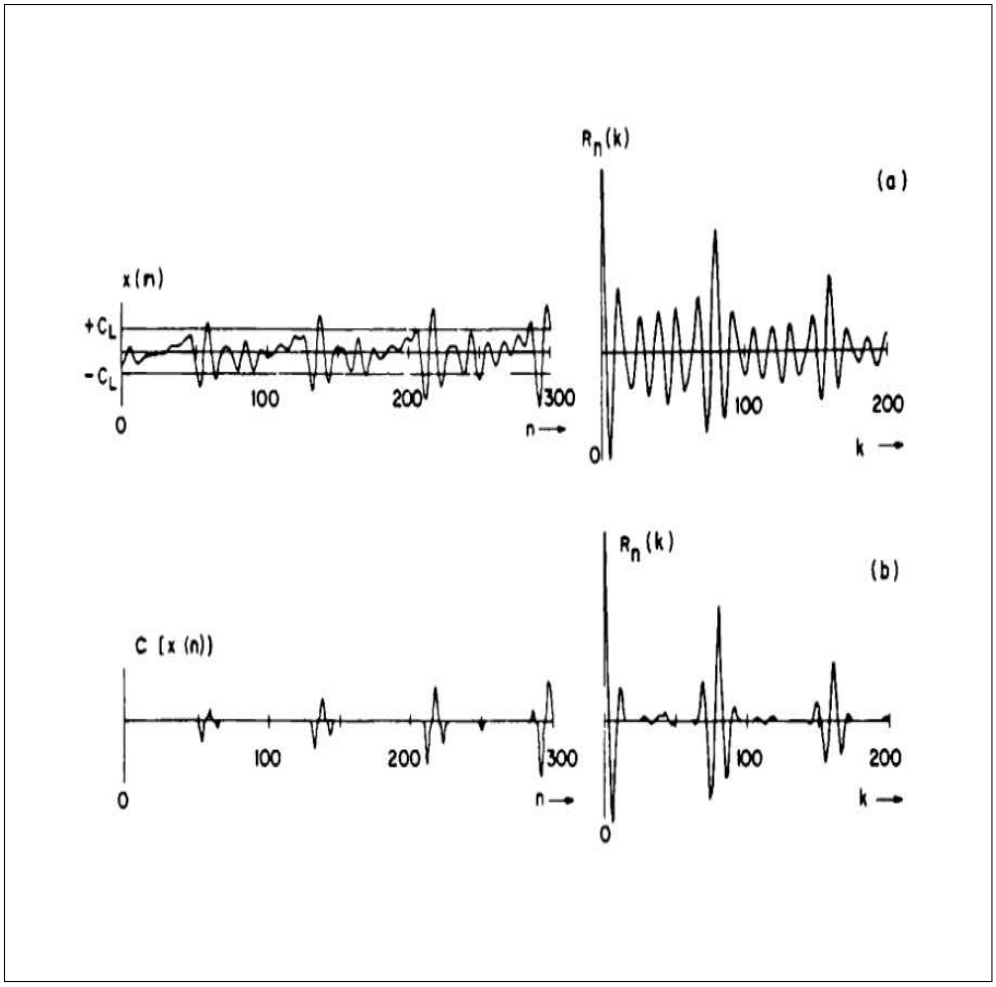

Figura 3.7: (a) Sonido vocalizado y su autocorrelacion. (b) sonido vocalizado recortado y su función de autocorrelación

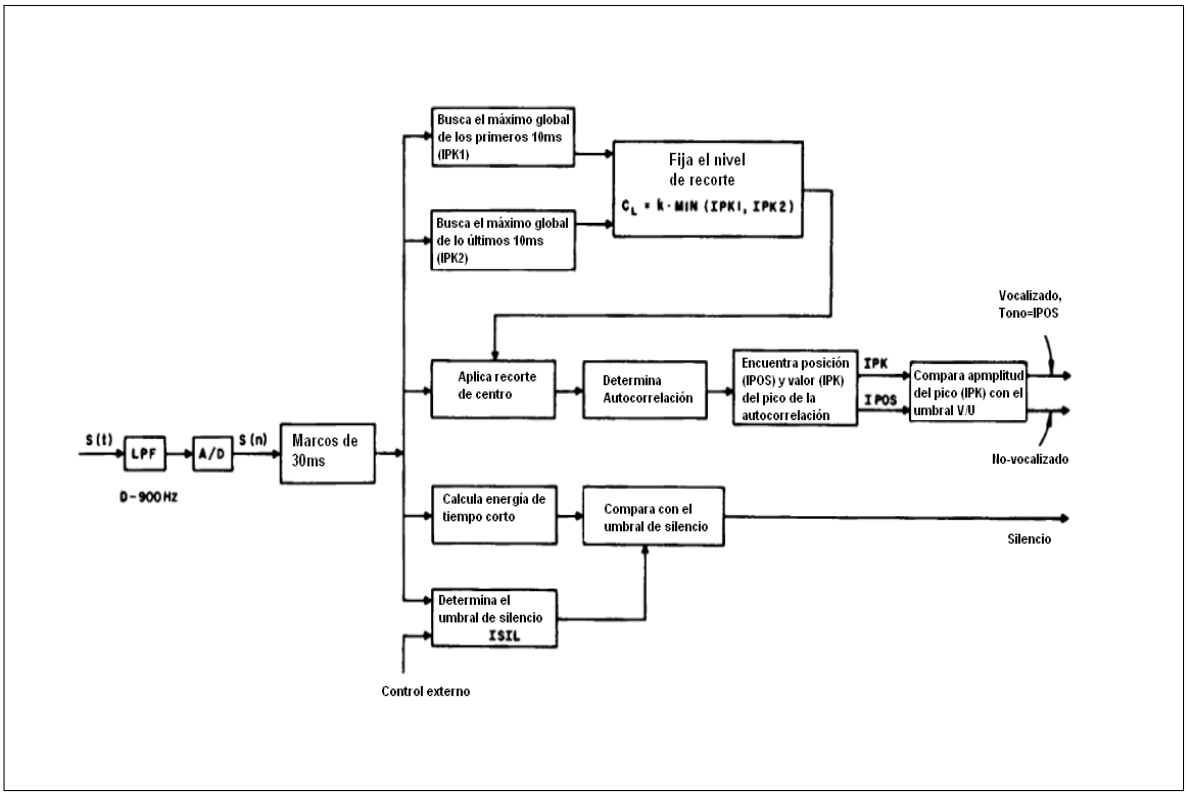

Figura 3.8: Algoritmo para discriminar silencio de voz, sonidos vocalizados de no-vocalizados y determinar el tono

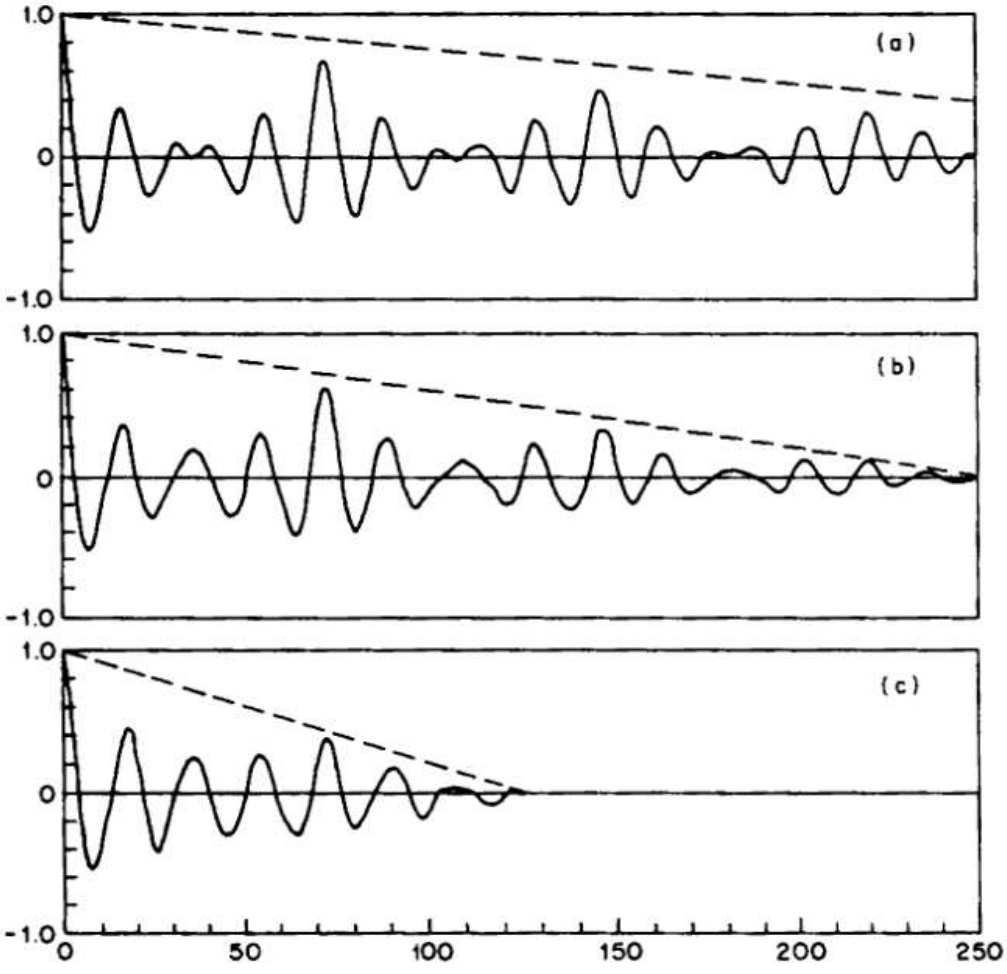

Figura 3.9: Efecto en la Autocorrelación de tempo corto debido a que al aumentar k se involucran menos datos en el cálculo de $R_n(k)$

$$\hat{R}_n(k) = \sum_{m=-\infty}^{\infty} x(m)w_1(n-m)x(m+k)w_2(n-k-m) \tag{3.22}$$

La ventana w_2 es una ventana mas grande que w_1 para permitir que se tome en cuenta mustras que quedan fuera del intervalo definido por w_1, para la ventana rectangular w_1 y w_2 se definen como:

$$w_1(m) = \begin{cases} 1 & 0 \leq m \leq N-1 \\ 0 & otro\ lugar \end{cases} \tag{3.23}$$

$$w_2(m) = \begin{cases} 1 & 0 \leq m \leq N-1+K \\ 0 & otro\ lugar \end{cases} \tag{3.24}$$

En lugar de desplazar la ventana hasta la posición m, se puede de manera equivalente dejar la ventana fija y desplazar la señal para que la posición m coincida con el origen (donde se ubica la ventana) de manera que la siguiente definición es igualmente válida:

$$\hat{R}_n(k) = \sum_{m=-\infty}^{\infty} x(n+m)w_1(m)x(n+m+k)w_2(m+k) \tag{3.25}$$

Para la ventana rectangular, esta definición es equivalente a:

$$\hat{R}_n(k) = \sum_{m=0}^{N-1} x(n+m)x(n+m+k) \quad \forall 0 \leq k \leq K \tag{3.26}$$

Estrictamente, la autocorrelación modificada es en realidad la correlación cruzada entre dos segmentos consecutivos de la señal, es decir entre el segmento $x(n+m)w_1(m)$ y el segmento $x(n+m)w_2(m)$

3.6.3. Algoritmo de Blankenship para el cálculo eficiente de la autocorrelación

Blankenship aprovecha el hecho de que la mayoría de los términos de la sumatoria para determinar $\hat{R}_n(k)$ aparecen dos veces, por ejemplo, para $k = 1$:

$$\hat{R}_n(1) = \sum_{m=0}^{N-1} x(m+n)x(m+n+1) \tag{3.27}$$
$$= x(n)x(n+1) + x(n+1)x(n+2) + ... + x(n+N-1)x(n+N) \tag{3.28}$$
$$= x(n+1)[x(n) + x(n+2)] + x(n+3)[x(n+2) + x(n+4)] + \tag{3.29}$$

De esta manera, el número de multiplicaciones se reduce a la mitad. El método de Blankenship expresa la autocorrelación como:

$$\hat{R}(k) = B(k) + C(k) \tag{3.30}$$

El tamaño de la ventana N se descompone en un componente par $(2qk)$m un componente opcional (ak) y un residuo (b).

$$N = 2qk + ak + b \tag{3.31}$$

donde $a = 0$, o bien $a = 1$ y b está en el rango $0 \leq b < k$.

$$B(k) = \sum_{j=0}^{q-1} \sum_{i=1}^{k} x(2jk + i + k)[x(2jk + i) + x(2jk + i + 2k)] \tag{3.32}$$

si $a = 0$:

$$C(k) = \sum_{i=1}^{b} x(2qk + i)x(2qk + i + k) \tag{3.33}$$

y si $a = 1$

$$C(k) = \sum_{i=1}^{b} x(2qk+i+k)[x(2qk+i+)+x(2qk+i+2k)]+ \sum_{i=b+1}^{k} x(2qk+i)x(2qk+i+k) \tag{3.34}$$

La siguiente función escrita en C calcula el k ésimo valor de la función de autocorrelación mediante el método de Blankenship

```
float blankenship(float x[],int n,int k) {
    int q,a,b,t,ti;
    float b_k=0,c_k=0,s;
```

```
    register int i,j;
    determina_q_a_b(k,n,&q,&a,&b);
    for (j=0;j<q;j++) {
       s=0;t=2*j*k;
       for (i=1;i<=k;i++) {
          ti=t+i;
          s+=x[ti+k]*(x[ti]+x[ti+2*k]);
       }
       b_k+=s;
    }
    t=2*q*k;
    if (a==0) {
       for (i=1;i<=b;i++) {
          ti=t+i;
          c_k+=x[ti]*x[ti+k];
       }
    }
    else {
       for (i=1;i<=b;i++) {
          ti=t+i;
          c_k+=x[ti+k]*(x[ti]+x[ti+2*k]);
       }
       for (i=b+1;i<=k;i++) {
          ti=t+i;
          c_k+=x[ti]*x[ti+k];
       }
    }
    return b_k+c_k;
}
```

La siguiente función escrita en C determina q, a y b en base a las siguientes restricciones $N = 2qk + ak + b$, $a = 0$ ó $a = 1$ y $0 \leq b < k$

```
void determina_q_a_b(int k, int n, int *q, int *a, int *b) {
    if (k) *q=n/(2*k);
    else fprintf(stderr,"determina_q_a_b: k VALE CERO!!!\n");
    if ((2*(*q)+1)*k <= n) *a=1; else *a=0;
    *b=n-2*(*q)*k-(*a)*k;
}
```

El método de Blankenship es particularmente útil cuando solo se requieren algunos valores de la autocorrelación, lo cual es el caso cuando se determinan los coeficientes LPC por el método de Durbin. En el caso de que se requieran todos los valores de la autocorrelación es mas recomendable hacerlo mediante la transformada discreta inversa de Fourier del espectro de energía (el cuadrado del espectro de magnitudes)

Capítulo 4

Procesamiento de la señal de voz en el dominio de la frecuencia

Hemos visto que procesando la señal en el dominio del tiempo podemos determinar si hay voz en la señal de audio, en caso de que haya voz, el tipo de sonido, es decir vocalizado o no-vocalizado y si se trata de sonidos vocalizados podemos determinar el tono. Sin embargo, el reconocimiento de voz, ya sea de palabras aisladas o de habla continua dificilmente se puede hacer en el dominio del tiempo, de hecho, el sistema auditivo humano realiza un análisis espectral de la señal de audio dado que dentro de la cloquea tenemos una colección de neuronas en forma de vellosidades que oscilan a frecuencias que dependen de la longitud de estas.

4.1. Transformada de Fourier de tiempo corto

No sirve de mucho obtener la transformada de toda la señal de voz capturada puesto que no sabríamos en que instante se presentaron los componentes de frecuencia que nos reporte la transformada de Fourier, si por ejemplo implementáramos un reconocedor de palabras aisladas no podríamos diferenciar carro de roca. Lo que necesitamos es hacer un análisis espectral pero de segmentos lo suficientemente cortos de audio como para que se pueda considerar que la señal en ese segmento corto es una señal estacionaria, es

decir, que los componentes de frecuencia no cambian significatívamente. Los segmentos no deben ser demasiado cortos tampoco, por regla general, deben ser lo suficientemente grandes como para que quepan dos periodos completos del componente de frecuencia mas baja que se esté considerando. A estos segmentos cortos se les conoce como *marcos*, es muy común utilizar marcos de 30ms para este propósito.

Aplicamos transformada de Fourier a la señal contenida en el marco (Ej de 30ms), hacemos avanzar el marco y volvemos a extraer la transformada de Fourier y así sucesívamente, de esta manera sabremos cuales son los componentes de frecuencia de la señal pero al desplazar el marco, sabremos también como van cambiando dichos componentes de frecuencia en el tiempo, en resumen, tendremos información de la señal en el dominio de la frecuencia pero sin perder el dominio del tiempo. Ante la pregunta de cuanto desplazar la ventana, diversos estudios indican que el oído humano no es capaz de percibir cambio bruscos realizados en menos de 10ms. Así como el ojo humano no es capaz de ver mas de 24 cuadros por segundo, el oído humano no puede escuchar frecuencias demasiado altas (arriba de 20KHz). Por lo aquí expuesto, es común avanzar los marcos en 10ms, si los marcos son de 30ms esto implica que el traslape entre marcos consecutivos es de dos tercios.

4.1.1. Aplicación de ventanas

Antes de determinar la transformada de Fourier de tiempo corto es conveniente aplicar una ventana que desvanezca la señal en los extremos del marco, en este sentido, se considera que aplicar la ventana rectangular es equivalente a no aplicar ninguna ventana. La ventana de Welch definida mediante (4.1) se muestra en la Figura 4.1. La ventana de Barlett definida mediante (4.2) se muestra en la Figura 4.2. La ventana de Hamming está definida mediante (4.3), la ventana de Hann está definida mediante (4.4), ambas forman parte de una familia de ventanas genéricas definidas mediante (4.5). La ventana de Hamming, la de Hann y la de Kaiser se muestran en la Figura 4.3.

$$welch(n) = 1 - \left(\frac{n - N/2}{N/2}\right)^2 \tag{4.1}$$

$$barlett(n) = \begin{cases} 1 - 2n/N & 0 \leq n \leq N/2 \\ 1 + 2n/N & -N/2 \leq n \leq 0 \\ 0 & otro\ lugar \end{cases} \tag{4.2}$$

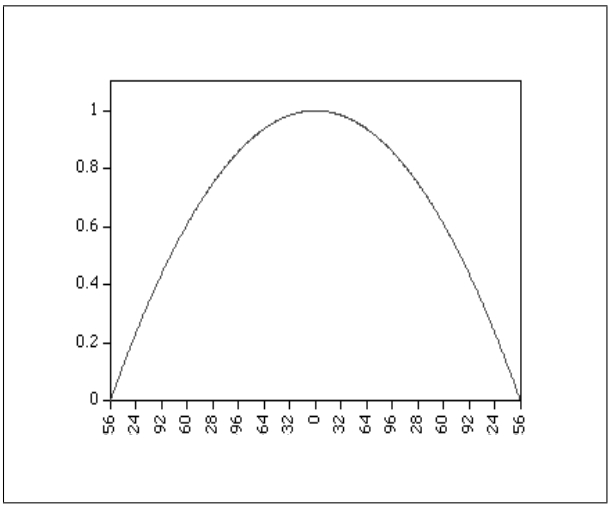

Figura 4.1: Ventana de Welch

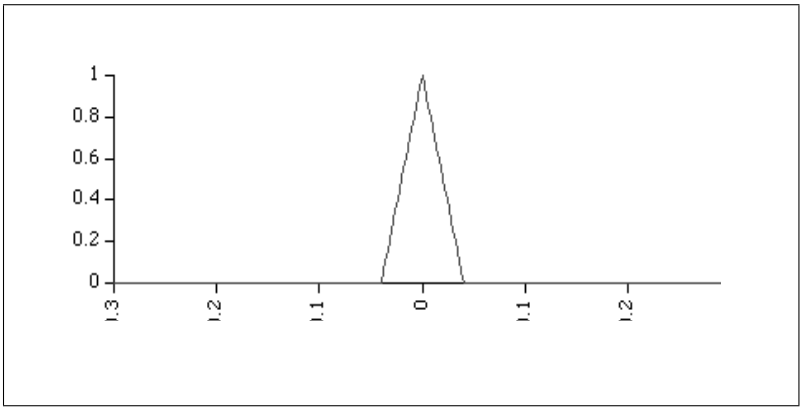

Figura 4.2: Ventana de Barlett

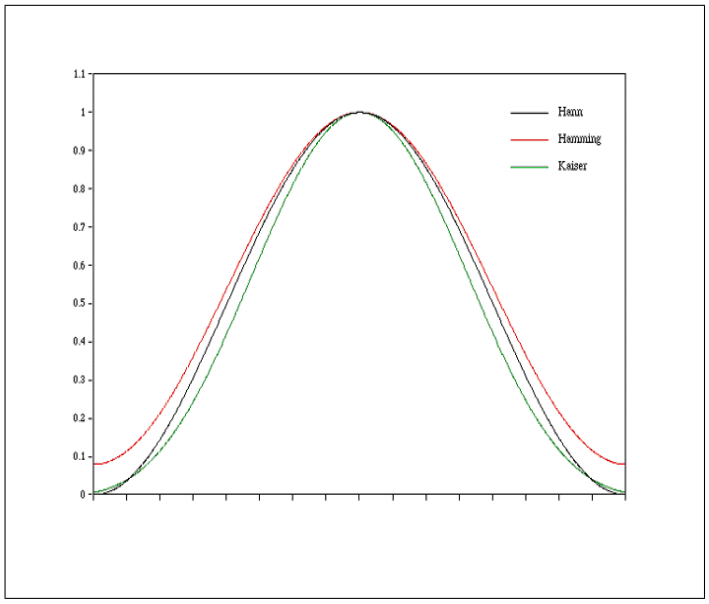

Figura 4.3: Ventanas de Hann, Hamming y Kaiser

$$hamming(n) = 0{,}54 + 0{,}46cos(2\pi n/N) \tag{4.3}$$

$$hann(n) = 0{,}5 + 0{,}5cos(2\pi n/N) \tag{4.4}$$

$$w(n) = a + (1 - a)cos(2\pi n/N) \tag{4.5}$$

4.1.2. El efecto "leakage"(escurrimiento)

Al tomar segmentos de tiempo corto de la señal de audio es altamente probable que dichos segmentos comiencen y/o terminen con un valor diferente de cero y además de valores distintos (entre el inicio y el final). La Transformada Discreta de Fourier entenderá esta diferencia entre el primer valor y el último como una discontinuidad, el cambio brusco que ve la Transformada de Fourier implicará un gran esfuerzo para reconstruirlo como una suma infinita de senoides lo cual se traduce en coeficientes de Fourier de alta frecuencia con alto contenido de energía. La energía total de la señal en el dominio de la frecuencia es igual a la energía de la señal en el dominio del

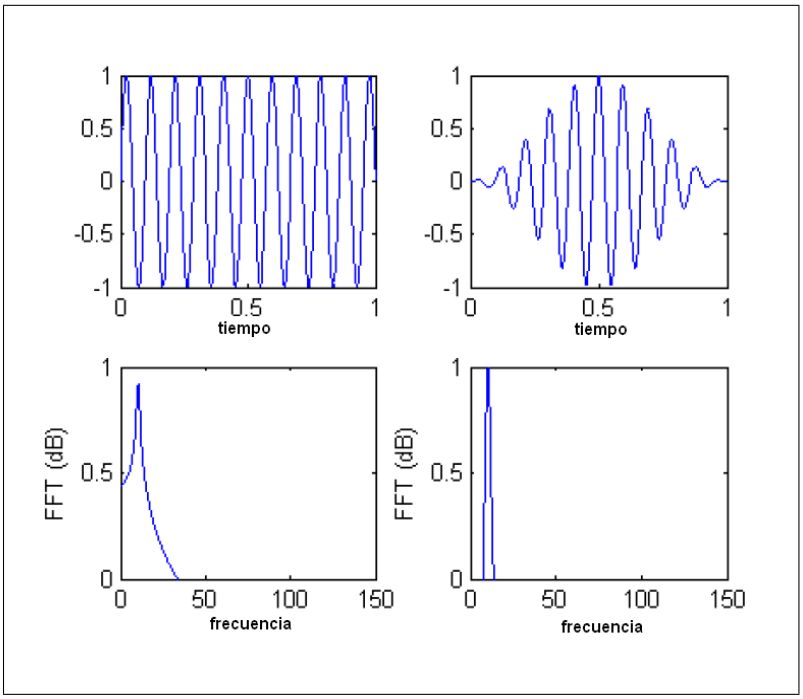

Figura 4.4: A la izquierda una senoide que no termina exactamente el el mismo valor con el que comienza y su espectro. A la derecha la aplicación de una ventana como la de Hamming corrige el problema

tiempo, eso es lo que nos dice el Teorema de Parseval, por lo tanto, la energía de los coeficientes de Fourier espurios de alta frecuencia será tomada de la energía de los coeficientes de Fourier verdaderos, a este efecto se le conoce como escurrimiento. Para reducir el efecto de escurrimiento se puede aplicar una ventana como las descritas arriba (de Hamming, de Parzen, etc), de esta manera tanto al inicio como al final del segmento de audio, la señal se desvanece y no existirá un escalón o cambio brusco en la señal cuando se le aplique transformada de Fourier. En la Figura 4.4 se muestra el espectro de frecuencias obtenido después de aplicar la transformada de Fourier a una senoide pura la cual no termina en el mismo valor con el que empieza, observe como parece que se *escurre* lo que debería ser un pico, en la misma figura se muestra la transformada de Fourier de la misma senoide después de aplicarle la ventana de Hamming

4.2. Determinación del espectrograma mediante la Transformada de Fourier de tiempo corto

No basta la información relativa a como se distribuye la energía en el dominio de la frecuencia al pronunciar una palabra para identificarla, es necesario saber de que forma los contenidos de frecuencia de la señal de voz evolucionan en el tiempo. Para convencernos de ello, pensemos en que si solo dispusieramos del dominio de la frecuencia no podríamos distinguir entre dos palabras con los mismos contenidos de frecuencia pero diferente distribución en el tiempo, por ejemplo no podríamos distinguir entre una elocución de la palabra *marca* y una elocución de la palabra *karma*.

Necesitamos tanto del dominio del tiempo como del dominio de la frecuencia, para ello, hacemos uso de la transformada de Fourier de tiempo corto, así, de cada marco de tiempo obtenemos su transformada de Fourier, previa aplicación de una ventana como la de Hamming, luego recorremos la ventana y repetimos el proceso, el resultado es una secuencia de espectros de frecuencia. Un espectrograma es una gráfica en la que el eje vertical corresponde con la frecuencia, el eje horizontal con el tiempo y la intensidad (puede ser nivel de gris o distribución de colores) corresponde con la amplitud de los coeficientes. Así pues, por cada marco de tiempo tendremos una posición fija en el eje horizontal y el espectro de magnitudes obtenido aplicando la transformada de Fourier de tiempo corto al marco en turno, se despliega en forma vertical variando la intensidad. Para apreciar de mejor manera el espectrograma se expresa la magnitud de los coeficientes en decibeles, es decir $20log_{10}(|X(k)|)$ en lugar de $|X(k)|$. En la Figura 4.5 se muestra el espectrograma de la frase *should i chase* en la parte superior se muestra la señal en el dominio del tiempo, abajo se muestra el espectrograma obtenido con marcos de 10ms y en la parte inferior se muestra el espectrograma obtenido con marcos de 40ms, se puede apreciar como al utilizar marcos de solo 10ms se logra una excelente resolución en el dominio del tiempo pero sacrificando la resolución en el dominio de la frecuencia, por otra parte, al ampliar los marcos a 40ms se logra que la transformada de Fourier capture de mejor manera los contenidos de frecuencia (particularmente de las frecuencias bajas) pero perdiendo resolución en el dominio del tiempo

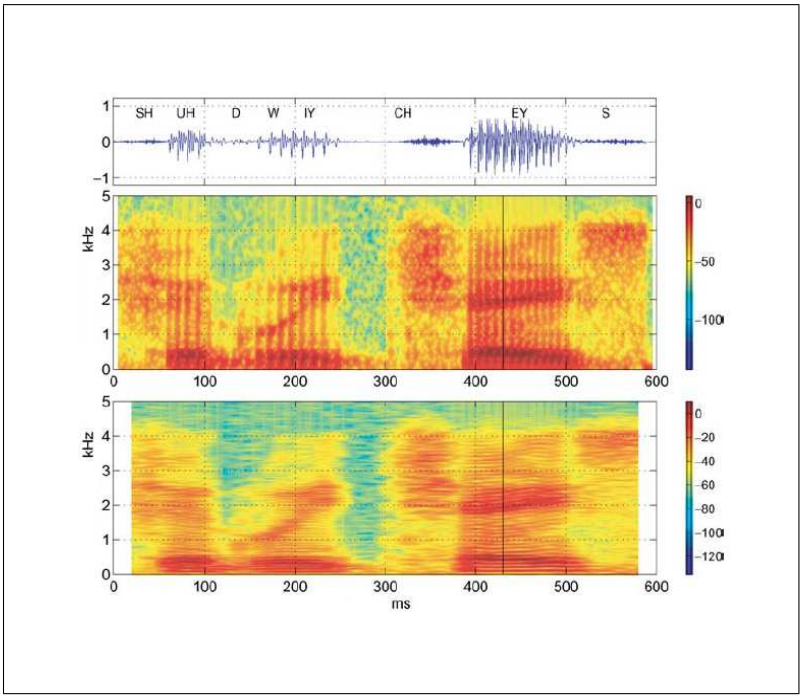

Figura 4.5: Espectrogramas determinados con marcos de 10ms (arriba) y 40 ms (abajo)

4.3. Determinación del espectrograma de la señal de voz mediante Bancos de Filtros

El espectrograma es un diagrama donde se despliega la cantidad de energía como función del tiempo y de la frecuencia simultáneamente, este diagrama tiene una resolución en el tiempo que depende del ancho de los marcos así como del traslape entre los marcos, mientras que la resolución en la frecuencia depende del ancho de los marcos´y de la frecuencia de muestreo.

Un espectrograma con resolución en frecuencia considerablemente menor se puede obtener mediante un banco de filtros, a la salida de cada filtro se determina la energía de tiempo corto en el dominio del tiempo, de esa manera la resolución en frecuencia depende del número de filtros que conforman el banco, si por ejemplo se utiliza un filtro pasabanda por cada banda crítica de Bark (con ancho de banda de un Bark), usaríamos 18 filtros y la resolución en frecuencia sería de 18 números por cada marco para construir el espectrograma de baja resolución.

El espectrograma obtenido de esta manera presenta muchas ventajas, primero que nada es fácil de implementar en hardware, es paralelizable y se puede usar como característica fundamental para reconocimiento de voz, el espectrograma de alta resolución es un arreglo con demasiados valores, resulta demasiado costoso para almacenar y casi prohibitiva la comparación entre espectrogramas de dos elocuciones, en este sentido el espectrograma de baja resolución es mucho mas deseable. A continuación se describirá a detalle la escala de Bark.

4.4. Escala de Bark

Es conveniente reproducir la manera en la que las personas identifican los sonidos, no todas las frecuencias se perciben con la misma sensibilidad, el oído humano percibe mejor las frecuencias bajas que las altas. La escala Bark fue diseñada para modelar el oído humano, es una escala psicoacústica que fué propuesta por Eberhard Zwicker en 1961 en honor a Heinrich Barkhausen quien realizó las primeras mediciones subjetivas de la percepción de la intensidad sonora, también conocida como sensación sonora o *sonoridad* (En Inglés *loudness*). La escala de Bark define 25 bandas *críticas*, cada banda crítica corresponde con un segmento de la cloquea, en la Tabla 4.1 se describen dichas bandas, cada banda crítica tiene un ancho de un Bark, para

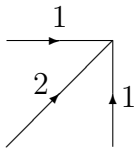

Figura 4.6: Restricción local simétrica de primer orden

convertir Hertz a Barks podemos utilizar la ecuación (4.6).

$$z = 13tan^{-1}\Big(\frac{0{,}76f}{1000}\Big) + 3{,}5tan^{-1}\Big(\frac{f}{750}\Big)^2 \tag{4.6}$$

donde z es frecuencia en Barks y f es la frecuencia en Hertz.

4.5. Doblado Dinámico en Tiempo

Alinear dos series de tiempos $R(n),\ 0 \leq n \leq N$ and $T(m),\ 0 \leq m \leq M$ es equivalente a encontrar una función de doblado $m = w(n)$ que mapea cada índice n de la serie R en un índice m de la serie T de manera que se realice un registro entre las dos series de tiempo. La función w está sujeta a las condiciones de frontera $w(0) = 0$ y $w(N) = M$ y a restricciones locales. Posíblemente la restricción local más utilizada es la que nos indica que si la trayectoria función óptima pasa por el punto (n, m), entonces debió pasar por $(n-1, m-1)$, por $(n, m-1)$ o por $(n-1, m)$ como se muestra en la Figura 4.6. Una penalización de 2 es impuesta cuando se elige $(n-1, m-1)$ y de 1 si se eligen $(n, m-1)$ o $(n-1, m)$, de esta manera, las tres posibles trayectorias de $(n-1, m-1)$ a (n, m) (Ej. ir primero a $(n, m-1)$ y luego a (n, m)) tendran todas el mismo costo de 2. Otras restricciones locales definidas por Itakura pueden usarse.

Sea $d_{n,m}$ la distancia entre el vector de características correspondiente al marco n de la elocución R y el vector de características correspondiente al marco m de la elocución T, entonces la función óptima entre R y T es aquella que minimiza la distancia acumulada $D_{n,m}$ definida mediante (4.7).

$$D_{n,m} = \sum_{p=1}^{n} d_{R(p),T(w(p))} \tag{4.7}$$

Tabla 4.1: La escala de Bark (las 25 bandas críticas)

Frec. inic (Barks)	Frec inic (Hz)	Frec final (Hz)
0	0	100
1	100	200
2	200	300
3	300	400
4	400	510
5	510	630
6	630	770
7	770	920
8	920	1080
9	1080	1270
10	1270	1480
11	1480	1720
12	1720	2000
13	2000	2320
14	2320	2700
15	2700	3150
16	3150	3700
17	3700	4400
18	4400	5300
19	5300	6400
20	6400	7700
21	7700	9500
22	9500	12000
23	12000	15500
24	15500	20000

Ya elegida una restricción local, $D_{N,M}$ puede calcularse utilizando la recurrencia definida por las ecuaciones (4.9),(4.10) y(4.11), la cual corresponde a la restricción local mostrada en la Figura 4.6. Basandose en esta recurrencia, $D_{N,M}$ se puede determinar utilizando programación dinámica.

$$D_{0,0} = d_{0,0} \tag{4.8}$$

$$D_{i,0} = d_{i,0} + D_{i-1,0} \tag{4.9}$$

$$D_{0,j} = d_{0,j} + D_{0,j-1} \tag{4.10}$$

$$D_{i,j} = min \begin{cases} D_{i-1,j-1} + 2d_{i,j} \\ D_{i-1,j} + d_{i,j} \\ D_{i,j-1} + d_{i,j} \end{cases} \tag{4.11}$$

En la Figura 4.7 se muestra la matriz de distancias que debemos llenar para implementar el doblado dinámico en tiempo mediante programación dinámica, en la misma figura se muestra la trayectoria óptima para llegar de la localidad (1,1) a la localidad (M,N) y que corresponde con la forma de la función de doblado óptima para alinear las series de tiempo de longitud N y M respectivamente que se muestran en la misma figura. La distancia entre las dos series de tiempo será la última en determinarse al llenar la matriz, es decir $D_{N,M}$, conviene normalizar esta distancia, es decir usar $D_{N,M}/(N+M)$, de esta manera, la distancia entre dos elocuciones de la misma palabra no será grande solo porque la palabra es larga, dicho de otra manera, la distancia entre dos elocuciones no debería ser pequeña solo porque la palabra es corta.

4.6. Distancias

Para poder determinar la distancia entre dos matrices con el mismo número de renglones y diferente número de columnas podemos utilizar el doblado dinámico en tiempo, pero para ello debemos ser capaces de determinar que tan diferente es cualquier columna de una matriz de cualquier columna de la otra matriz, estas columnas se pueden ver como vectores del mismo tamaño (numero de componentes), las siguientes distancias sirven para comparar vectores:

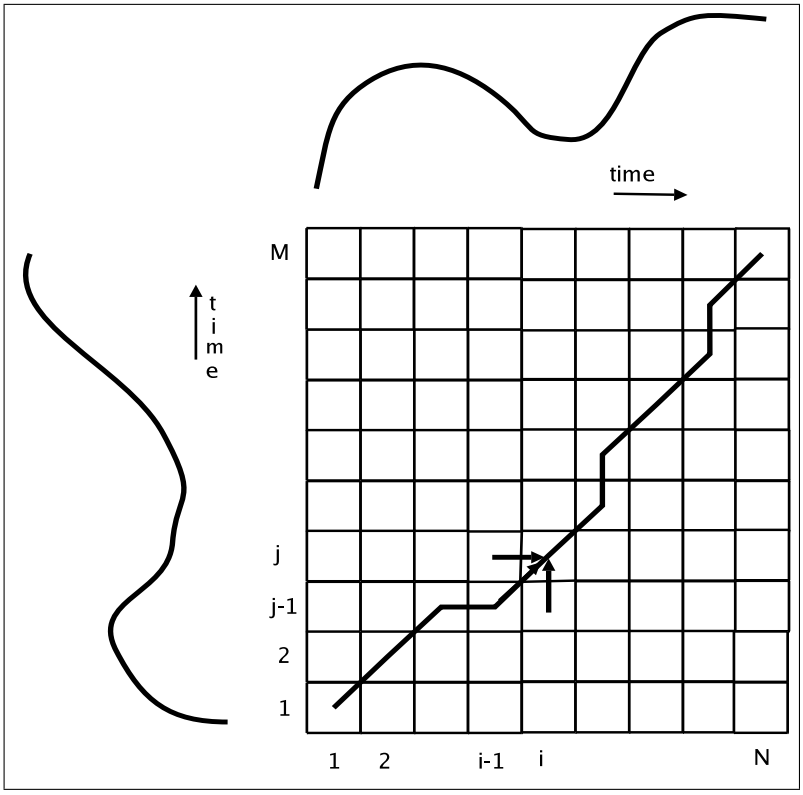

Figura 4.7: Matriz de distancias para implementar doblado dinámico en tiempo y la trayectoria óptima de acuerdo a la restricción local simétrica de primer orden

4.6.1. Distancia de Manhatan

De acuerdo a esta distancia, los puntos que están a lo largo de la periferia de un cuadrado están a la misma distancia del centro, la distancia de Manhatan recibe este nombre debido a que en una ciudad donde las calles son rectas (como en Manhatan) la distancia entre dos lugares de la misma ciudad la medimos en cuadras pues no podemos ir en línea recta.

$$d(a,b) = \sum_{i=1}^{d} |a_i - b_i| \tag{4.12}$$

4.6.2. Distancia Euclidiana

De acuerdo a esta distancia, los puntos que están en la periferia de un círculo están a la misma distancia del centro, es quizás la distancia mas fácil de concebir puesto que tenemos el concepto de que la distancia mas corta entre dos puntos es la línea recta y podemos deducir esta distancia utilizando el Teorema de Pitágoras

$$d(a,b) = \sqrt{\sum_{i=1}^{d} (a_i - b_i)^2} \tag{4.13}$$

4.6.3. Distancias LP

También conocida como distancias de Minkowski es la generalización de las distancias eucludiana y de Manhatan, aa distancia Euclidiana es por ello conocida como la distancia L2, es decir es la distancia de Minkowsky cuando $p = 2$. La distancia de Manhatan es conocida también como la distancia L1 ya que es la distancia de Minkowski cuando $p = 1$.

$$d(a,b) = \Big(\sum_{i=1}^{d} (a_i - b_i)^p \Big)^{1/p} \tag{4.14}$$

4.6.4. Distancia coseno

Al comparar vectores, a veces es conveniente que la comparación se haga de manera que esta no dependa de la magnitud de los vectores bajo comparación, por ejemplo, si un vector de características se forma determinando

la energía contenida en cada banda de frecuencia, entonces la magnitud de tal vector dependerá del volumen con el que se grabó y no del fonema o tipo de sonido grabado del que se extrajo dicho vector de características. La distancia coseno entre dos vectores depende de la orientación de estos y no de su magnitud, básicamente se trata de determinar en ángulo menor que hay entre los dos vectores bajo comparación, el producto punto entre dos vectores se determina mediante:

$$a \cdot b = |a||b|cos\theta \tag{4.15}$$

donde θ es el ángulo menor que hay entre los vectores a y b

$$cos\theta = \frac{A \cdot B}{|a||b|} = \frac{a_1b_1 + a_2b_2 + ... + a_db_d}{\sqrt{a_1^2 + a_2^2 + ... + a_d^2}\sqrt{b_1^2 + b_2^2 + ... + b_d^2}} \tag{4.16}$$

$$cos\theta = \frac{\sum_{i=1}^{d} a_ib_i}{\sqrt{\sum_{i=1}^{2} a_i^2}\sqrt{\sum_{i=1}^{d} b_i^2}} \tag{4.17}$$

Esta es en realidad una medida de similitud no una distancia, es decir mientras menos se parecen los vectores la medida de similitud se aproxima mas a cero, como el mayor valor que el coseno devuelve es 1.0, podemos convertir la medida de similitud en distancia de la siguiente manera:

$$d(a,b) = 1 - |cos\theta| = 1 - \left| \frac{\sum_{i=1}^{d} a_ib_i}{\sqrt{\sum_{i=1}^{d} a_i^2}\sqrt{\sum_{i=1}^{d} b_i^2}} \right| \tag{4.18}$$

4.7. Criterio del vecino más cercano y criterio de los K-vecinos

Para identificar un patron al que llamaremos *consulta* se puede buscar de entre los patrones que conforman el diccionario, es decir de entre la colección de patrones previamente etiquetados y declarar que la consulta pertenece a la clase del patrón que se encuentre a menor distancia, siempre y cuando esa distancia sea menor que cierto umbral, si la distancia al patrón de referencia mas cercano es mayor que dicho umbral se debe declarar que no es posible identificar a la consulta. El criterio consistente en etiquetar a la consulta copiando la etiqueta del patrón de referencia que se encuentre a menor

distancia se conoce como *criterio del vecino más cercano*. Buscar al más cercano no necesariamente debe hacerse de manera secuencial, existen índices de proximidad como los basados en BK-trees, Fixed Query trees y muchos otros que son capaces de encontrar al patron mas cercano comparando solo con algunos patrones del diccionario denominados *pivotes*.

Si en lugar de buscar solo al patrón mas cercano a la consulta buscamos a los K patrones mas cercanos a la consulta, entonces estamos en posibilidad de usar el criterio de los K-vecinos que es un esquema de votación donde solo votan los K patrones mas cercanos a la consulta. La consulta se declara como perteneciente a la clase a la que pertenecen los patrones que ganen la votación. Se recomienda que K sea un número impar para disminuir la probabilidad de empates.

El esquema basado en K-vecinos disminuye el régimen de errores de clasificación, sin embargo requiere que el diccionario tenga por cada clase mas de un patron lo cual complica la implementación, por ejemplo implica colectar por cada palabra del diccionario no solo una sino varias elocuciones, esto también incrementará el tamaño de la colección de patrones complicando las búsquedas.

4.8. Implementación de un sistema de reconocimiento de palabras aisladas mediante espectrogramas

Los ingenieros que trabajaban en los Laboratorios Bell de la At&T imprimían los espectrogramas de un conjunto reducido de palabras con las que hacían pruebas, entonces se dieron cuenta de que eran capaces de identificar una palabra con solo ver su espectrograma, esto significaba trasladar un problema de identificación de una señal de audio a un problema de reconocimiento de imágenes, si bien esto podría significar una herramienta para los sordos (que no sean ciegos), el hecho es que los espectrogramas se pueden utilizar como una caracterización de la señal de voz para realizar reconocimiento de palabras aisladas. Sin embargo, un espectrograma tiene demasiada información, por ejemplo si una señal de voz es muestreada a 8000 Hz y los marcos se conforman de 256 muestras (una potencia de 2), lo cual corresponde aproximadamente a 30 ms, entonces el espectrograma sería un arreglo de 128 coeficientes (recuerde la simetría del espectro de magni-

tudes de una señal real) por cada marco. Si los marcos se avanzan 10 ms cada vez, entonces el espectrograma de una palabra de medio segundo sería una matriz de 128 por 50, es decir, 6400 coeficientes. Para implementar un sistema de reconocimiento de palabras aisladas conviene realizar una extracción de características que entregue pocos coeficientes facilitando los demás procesos del sistema de reconocimiento de voz.

Del espectrograma se puede determinar la energía por cada banda crítica de Bark. Si se utilizan de la banda crítica de Bark 1 a la 16 (excluyendo a la banda 0) obtendríamos un arreglo de 16x50, es decir 800 valores en lugar de los 6400 descritos anteriormente. Para determinar la energía contenida en una banda crítica de algún marco simplemente se suman los cuadrados de las magnitudes de los coeficientes espectrales correspondientes a dicha banda:

$$E = \sum_{k=k1}^{k2} |X(k)|^2 \tag{4.19}$$

donde $k1$ es el índice del coeficiente espectral correspondiente al inicio de la banda y $k2$ es el índice del coeficiente espectral correspondiente al final de la banda en cuestión.

Con este sistema de extracción de características se puede implementar un programa que por cada palabra pronunciada ante el micrófono, agregue la entrada correspondiente al diccionario del sistema de reconocimiento de voz, tal diccionario es un archivo con las etiquetas y características de las palabras conocidas por el sistema, tal diccionario puede tener el siguiente formato:

```
<Etiqueta> <Num de marcos>

1,<Energía Banda 1>,<Energía Banda 2>, ...,<Energía Banda 16>
2,<Energía Banda 1>,<Energía Banda 2>, ...,<Energía Banda 16>

:

<Etiqueta> <Num de marcos>

1,<Energía Banda 1>,<Energía Banda 2>, ...,<Energía Banda 16>
2,<Energía Banda 1>,<Energía Banda 2>, ...,<Energía Banda 16>

:
```

El sistema de reconocimiento procede de la misma manera que el programa con el que se creó el diccionario en lo concerniente a la extracción de características, excepto que en lugar de agregar otra entrada al diccionario lo recorre comparando cada matriz de energías de la palabra recién capturada con cada matriz de energías de las palabras contenidas en el diccionario utilizando doblado dinámico en tiempo para realizar la comparación y detectando aquella con la que la distancia es menor, si esta distancia es menor que cierto valor umbral, entonces reportar ocurrencia de la palabra cuya distancia es menor leyendo su etiqueta del diccionario para tal efecto, si la distancia es mayor al umbral declarar que la palabra es desconocida para el sistema.

4.9. Evaluación de la transformada continua de Fourier en los ceros de los polinomios de Hermite

La Transformada discreta de Fourier (DFT por sus siglas en inglés) tiene ciertas restricciones, no funciona adecuadamente cuando la señal no es estacionaria o tiene estructural fractal. La Transformada Continua de Fourier (CFT por sus siglas en inglés) no tiene semejante restricción, normalmente se considera que la CFT no se puede evaluar pues tiene un kernel infinito, sin embargo, una discretización de la CFT fue propuesta en [6]. Los pasos necesarios para determinar la discretización de la CFT son:

1. Convertir la secuencia 0,1,2,...,N-1 a N valores equiespaciados desde $-\pi$ hasta π

2. Ajustar el segmento de la señal de voz a un polinomio trigonométrico de grado $M < N/2$ dado por (4.20) usando las fórmulas (4.21) y (4.22) para encontrar a_j and b_j respectivamente [7].

$$\frac{a_0}{2} + \sum_{j=1}^{M} [a_j cos(jx) + b_j sin(jx)] \tag{4.20}$$

$$a_j = \frac{2}{N} \sum_{k=1}^{N} [f(x_k) cos(jx_k)] \quad \forall \quad j = 0, 1, \ldots, M \tag{4.21}$$

$$b_j = \frac{2}{N}\sum_{k=1}^{N}[f(x_k)sin(jx_k)] \quad \forall \quad j = 1, 2, \ldots, M \tag{4.22}$$

3. Forma el vector x con las raíces del polinomio de Hermite de grado P

4. Construya el Kernel de Fourier (matriz F) [8] usando la ecuación (4.23). Esta matriz es Hermitiana, de modo que $F^{-1} = F^t$. multiplicar por F es equivalente a encontrar una discretización de la CFT, multiplicando por su transpuesta encontramos una discretización de la Transformada de Fourier inversa.

$$F_{i,j} = \frac{\pi}{\sqrt{2n}}\sqrt[4]{\frac{4n+3-x_j^2}{4n+3-x_i^2}}[cos(x_i x_j) + jsin(x_i x_j)] \tag{4.23}$$

5. Evaluar el polinomio trigonométrico encontrado en el paso (2) en el vector de ceros de Hermite determinado en el paso (3), llamemos a este vector f.

6. Determinar $g = Ff$, g es la discretización de la CFT de f

Conviene utilizar las raíces equiespaciadas que se encuentran en la parte central de un polinomio de Hermite de alto grado, por ejemplo, las 170 raíces centrales de un polinomio de Hermite de grado 480.

Capítulo 5

Procesamiento Homomórfico de la señal de voz

Como hemos visto, la señal de voz puede considerarse como el resultado de la convolución entre un vector con los coeficientes de un filtro que modela al tracto vocal y otro vector con un tren de impulsos o bien con un vector con números aleatorios que modelan la excitación del tracto vocal de sonidos vocalizados y no vocalizados respectivamente. El análisis de la señal de voz puede entenderse como el problema de determinar los componentes de una convolución partiendo del resultado de la convolución, es decir de la señal de voz. Al proceso de separar los componentes de una convolución le llamamos *deconvolución*. La deconvolución homomórfica se lleva a cabo mediante un concepto denominado filtrado homomórfico.

Un sistema homomórfico para la convolución obedece un principio generalizado de superposición (donde la operación suma es reemplazada por la operación convolución) que dicta que si la entrada se compone de una combinación lineal de señales elementales, entonces la salida será una combinación lineal de las salidas que produciría dicho sistema para las señales elementales si cada una es aplicada al sistema de manera independiente. Se define pues a la clase de sistemas que obedecen el principio generalizado de superposición como sistemas homomórficos para la convolución. En la Figura 5.1 se muestra la representación de un sistema homomórfico para la convolución.

Todo sistema homomorfico puede descomponerse en tres sistemas homomorficos en cascada, el primero toma entradas combinadas por convolución y las transforma en una combinación aditiva de las salidas correspondientes. El segundo sistema es un sistema lineal convencional que obedece el princi-

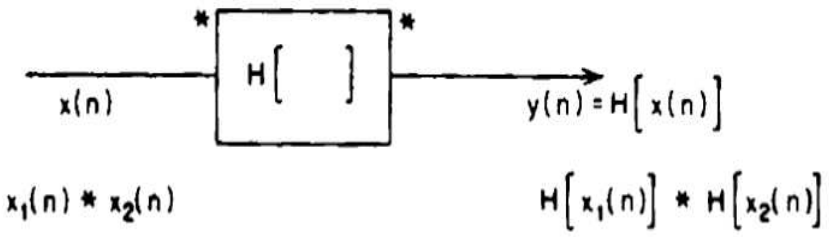

Figura 5.1: Representación de un Sistema Homomórfico para la convolución

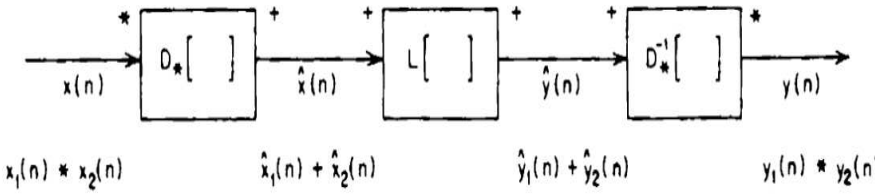

Figura 5.2: Forma canónica de un sistema para deconvolución homomórfica

pio de superposición tradicional. El tercer sistema es el inverso del primero, entonces transforma señales combinadas aditivamente en señales combinadas por convolución. La descomposición del sistema homomórfico en los tres sistemas descritos se conoce como conversión del sistema a su forma canónica. En la Figura 5.2 se muestran los tres sistemas, al sistema $D_*[]$ se le denomina *Sistema característico para deconvolución homomórfica.*

La Transformada Z de la salida de un sistema característico debe ser una combinación aditiva de transformadas Z, entonces el comportamiento en el dominio de la frecuencia de un sistema característico para convolución debe tener la propiedad de que si la señal está representada como un producto de transformadas Z a la entrada, entonces la salida debe ser una suma de las transformadas Z correspondientes. Un enfoque para realizar tal sistema se

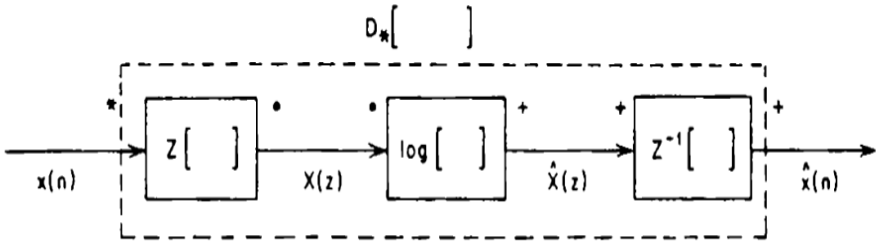

Figura 5.3: Representación del sistema característico para deconvolución homomórfica

basa en el hecho de que el logaritmo de un producto es igual a la suma de los logaritmos. El sistema característico puede entonces representarse como en la Figura 5.3.

5.1. El Cepstrum y el Cepstrum complejo

La Transformada inversa de Fourier del logaritmo complejo de la transformada de Fourier de la entrada es la salida de un sistema característico para convolución. La salida del sistema característico se conoce como *Cepstrum complejo* pues se utiliza el logaritmo complejo. El término *Cepstrum* denota a la Transformada inversa de Fourier del logaritmo de la magnitud de la transformada de Fourier de una señal. en la Figura 5.4 se muestra un diagrama de la determinación del Cepstrum y del Cepstrum complejo.

5.2. Aplicación a la estimación del tono

El cepstrum permite distinguir entre sonidos vocalizados y no-vocalizados, además, el tono se puede determinar a partir del cepstrum. Para sonidos vocalizados ocurre un pico en el cepstrum ubicado en una posición que coincide con el periodo de la frecuencia fundamental de la señal de voz, este pico no ocurre para segmentos de sonido no-vocalizado. Estas propiedades del cepstrum se pueden aprovechar para determinar si un segmento de voz es

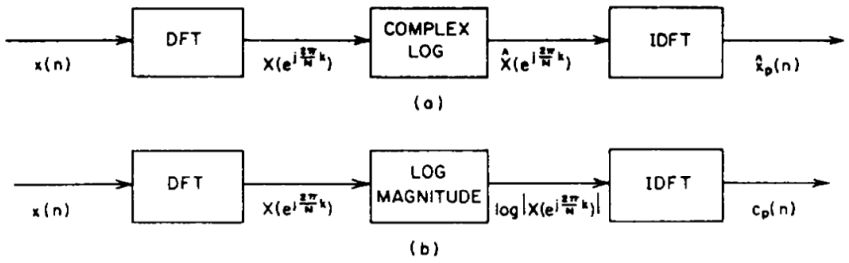

Figura 5.4: (a) El Cepstrum complejo y (b) El Cepstrum

vocalizado y si lo es, estimar el periodo fundamental, es decir, el tono.

En el cepstrum buscamos un pico en la vecindad del periodo del tono esperado. Si el pico es superior a un umbral pre-establecido, el segmento se declara como vocalizado y la posición del pico es una buena estimación del tono, si el pico no supera el valor umbral, el segmento de voz se declara como no-vocalizado, en vista de que los parámetros de excitación de la señal de voz no cambian muy rápidamente, el análisis se puede repetir cada 10-20 ms. En la Figura 5.5 se muestra una serie de log-spectra y su cepstra correspondiente para un parlante masculino, los primeros siete segmentos son no-vocalizados y los restantes son vocalizados con un periodo del tono que se va incrementando al transcurrir el tiempo.

5.3. Aplicación a la estimación de los formantes

Podemos asumir que la parte inferior del cepstrum corresponde principalmente al tracto vocal, al pulso glotal y la radiación de la voz mientras que la parte superior corresponde a la excitación. Los picos del espectro corresponden a las frecuencias formantes, esto sugiere que los formantes pueden ser estimados localizando los picos en el log-espectro *suavizado cepstralmente*.

Los picos del log-espectro son localizados y se decide si el sonido es vocalizado o no a partir del cepstrum. Si el segmento de voz es vocalizado, el tono se estima a partir del cepstrum y los primeros tres formantes son estimados con los picos del log-espectro. En el caso de que el sonido sea no-vocalizado se busca el pico mas alto del log-espectro y se asume que ahí se ubica un polo,

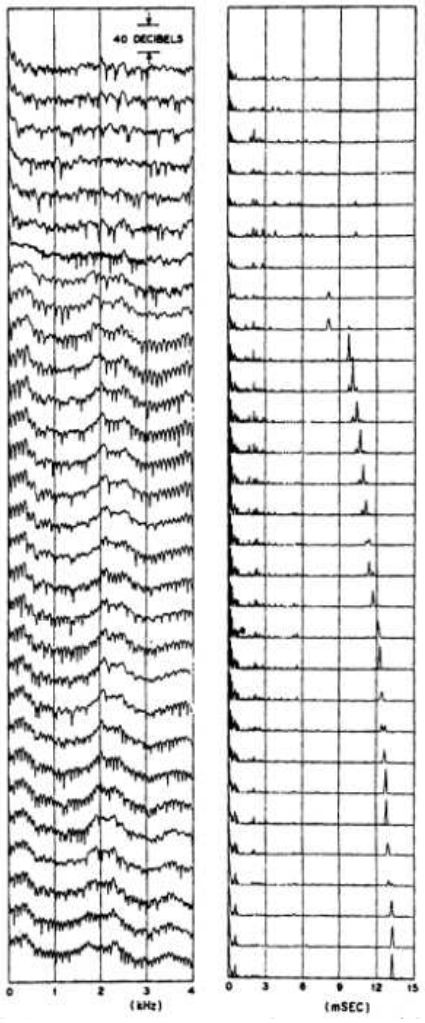

Figura 5.5: Un pico ubicado en el periodo del tono ocurre en el cepstrum como una indicación de que el sonido es vocalizado, los primeros 7 segmentos no son vocalizados, el resto sí lo son y puede verse como el tono va aumentando su periodo gradualmente

también se busca un cero asumiendo que la diferencia entre el pico mas alto y el valle mas bajo coincide con el rango dinámico de la frecuencia contenida en la señal.

En la Figura 5.6 se ilustra una estimación del tono y de las frecuencias formantes, la parte izquierda muestra una secuencia de cepstra calculados a intervalos de 20 ms. A la derecha el log-espectro de magnitudes es graficado junto con su correspondiente log-espectro suavizado cepstralmente sobrepuesto. Las líneas que conectan los picos son seleccionados como los primeros tres formantes, se puede observar como dos frecuiencias formantes ocasionalmente se juntan demasiado al grado de que ya no corresponden a dos picos distintos, esta situación se puede resolver realizando análisis espectral mediante la transformada Z chirp.

5.4. Determinación de los coeficientes MFCC (Mel-frequency Cepstral coefficients)

Los coeficientes cepstrales de Mel se determinan de acuerdo a ISP (Intelligent Sound Implementation) [9]. Primero la señal de audio es dividida en marcos de tiempo por ejemplo de 30 ms, se aplica una ventana de Hamming a cada marco de tiempo y luego la transformada discreta de Fourier, es otras palabras se determina $X(k)$ mediante:

$$X(k) = \sum_{n=0}^{N-1} [x(n)w(n)]e^{-j2\pi kn/N} \quad k = 0, 1, ..., N-1 \tag{5.1}$$

donde N es el tamaño del marco en muestras, k corresponde a la frecuencia $f(k) = kf_s/N$ (f_s es la frecuencia de muestreo) y $w(n)$ es la ventana de Hamming dada por:

$$w(n) = 0{,}54 - 0{,}46cos(\pi n/N) \tag{5.2}$$

Enseguida, la magnitud del espectro $|X(k)|$ se escala logaritmicamente tanto en frecuencia como en magnitud, para escalar logaritmicamente en frecuencia se utiliza el banco de filtros de Mel $H(k, m)$ y luego se escala logaritmicamente la magnitud para obtener:

$$X'(m) = ln\left(\sum_{k=0}^{N-1} [|X(k)|H(k, m)]\right) \tag{5.3}$$

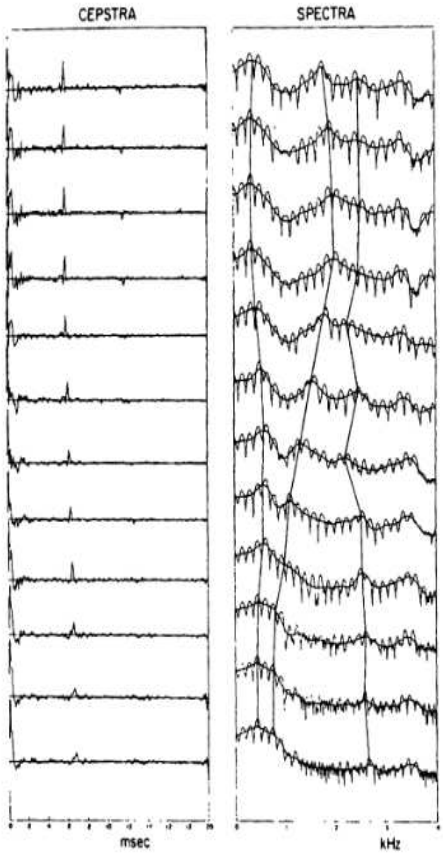

Figura 5.6: (Izquierda) Un pico ocurre en el cepstrum cuando se trata de sonidos vocalizados, el pico ocurre en la posición correspondiente al tono (Derecha) El log-espectro de magnitudes y el log-espectro suavizado cepstralmente permiten localizar los formantes

para $m = 1, 2, ..., M$, donde M es el número de filtros del banco, por supuesto $M << N$.

5.4.1. La escala de Mel

La Escala Mel, propuesta por Stevens, Volkman y Newmann en 1937, es una escala musical perceptual de tonos juzgados como intervalos equiespaciados por parte de observadores.

El punto de referencia entre esta escala y la frecuencia normal se define equiparando un tono de 1000 Hz, 40 dBs por encima del umbral de audición del oyente, con un tono de 1000 mels. Por encima de 500 Hz, los intervalos de frecuencia espaciados exponencialmente son percibidos como si estuvieran espaciados linealmente. En consecuencia, cuatro octavas en la escala de hercios por encima de 500 Hz se comprimen a alrededor de dos octavas en la escala mel.

Muchos músicos y psicólogos prefieren una representación bidimensional del tono mediante el color de tono (o croma) y altura de tono, o una representación tridimensional como la estructura helical propuesta por Roger Shepard, al representar más adecuadamente otras propiedades de la audición musical.

Para convertir f de Hertz a Mels se emplea:

$$\phi = 2595 log_{10}\Big(\frac{f}{700} + 1\Big) \tag{5.4}$$

que es una buena aproximación muy utilizada, en la Figura 5.7 se muestra la curva de Mels vs Hertz

5.4.2. Filtros triangulares de Mel

El banco de filtros de Mel es una colección de filtros triangulares definidos por sus frecuencias centrales $f_c(m)$

$$H(k,m) = \begin{cases} 0 & f(k) < f_c(m-1) \\ \frac{f(k)-f_c(m-1)}{f_c(m)-f_c(m-1)} & f_c(m-1) \leq f(k) < f_c(m) \\ \frac{f(k)-f_c(m+1)}{f_c(m)-f_c(m+1)} & f_c(m) \leq f(k) < f_c(m+1) \\ 0 & f(k) \geq f_c(m+1) \end{cases} \tag{5.5}$$

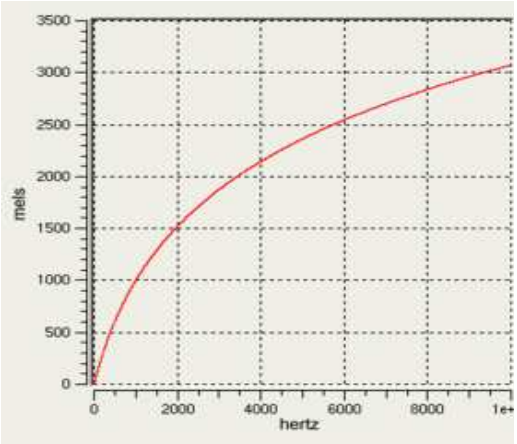

Figura 5.7: Escala de Mel

En la Figura 5.8se muestran los filtros triangulares de Mel de acuerdo a las especificaciones del HTK (Hidden Markov Model Kit)

Aunque en Hertz, las frecuencias centrales están repartidas logaritmicamente, en la escala de Mel las frecuencias centrales están linealmente distribuidas, la separación entre cada filtro triangular de Mel y el que le sigue está dada por:

$$\Delta\phi = \frac{\phi_{max} - \phi_{min}}{M+1} \tag{5.6}$$

donde ϕ_{max} y ϕ_{min} son los límites del rango de frecuencias de la señal bajo análisis correspondientes a f_{max} y f_{min} respectivamente y M es el número de filtros que conforman el banco de filtros.

Las frecuencias centrales de Mel están dadas por:

$$\phi_c(m) = m\Delta\phi \quad m = 1, 2, ..., M \tag{5.7}$$

Para obtener las frecuencias centrales en Hertz utilizamos la inversa de (5.4) y obtenemos:

$$f_c(m) = 700\Big(10^{\phi_c(m)/2595} - 1\Big) \tag{5.8}$$

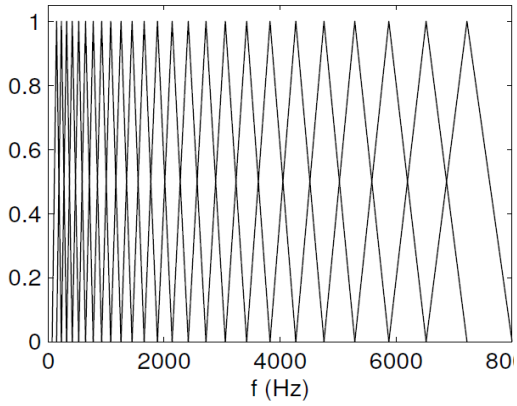

Figura 5.8: Filtros Triangulares de Mel

5.4.3. Transformada Coseno de la salida de los filtros de Mel

El último paso para obtener los coeficientes cepstrales de Mel es aplicar la transformada coseno discreta (DCT por sus siglas en inglés) mediante:

$$c(l) = \sum_{m=1}^{M} X'(m) cos(l\pi(m - \frac{1}{2})/M) \quad l = 1, 2, ..., M \tag{5.9}$$

$c(l)$ es el l-ésimo coeficiente cepstral de Mel.

5.5. Sistema de reconocimiento de voz usando los MFCC

El reconocedor de palabras aisladas en base a coeficientes cepstrales de Mel es prácticamente igual al sistema de reconocimiento de palabras aisladas mediante espectrogramas descrito en la sección 4.8. Ambos sistemas caracterizan una palabra mediante una matriz de flotantes, en ambos casos la matriz tiene un número fijo de columnas y un número variable de renglones pues este

depende de la duración de la elocución correspondiente. La única diferencia entre ambos sistemas es el denominado *front end*, es decir , el módulo de extracción de características de la señal de voz y por supuesto la interpretación que debe de hacerse de dicha matriz, en el caso del reconocedor mediante espectrogramas descrito en 4.8 cada columna corresponde con una banda crítica de Bark, mientras que en el reconocedor hecho en base a coeficientes LPC, cada columna corresponde con uno de los filtros triangulares de Mel. Los demás módulos no es entonces necesario cambiarlos, para comparar matrices se puede usar el doblado dinámico en tiempo, se pueda usar el mismo criterio del vecino mas cercano o de k-vecinos y usar distancia Euclidiana o distancia coseno. Para implementar un sistema de reconocimiento de voz en base a MFCC se propone utilizar quince filtros triangulares de Mel, es decir $M = 15$, una frecuencia inferior f_{min} de 80 Hertz puesto que es la frecuencia mínima que un tracto vocal puede emitir, y una frecuencia superior f_{max} de 4000 Hertz.

Capítulo 6

Codificación Lineal Predictiva

La codificación lineal predictiva se ha convertido en la técnica predominante de extracción de características de la señal de voz, presenta las siguientes ventajas:

1. Estimación mas precisa de los parámetros de voz, es decir, del espectro, los formantes, energía y área del tracto vocal.

2. Bajo régimen de bits, este se define como el número de bits por segundo que es necesario muestrear, por ejemplo un régimen de bits de 64000 bits/seg puede significar 8000 muestras de 8 bits cada segundo, o bien, 4000 muestras de 16 bits cada segundo.

3. Relativo bajo costo computacional

6.1. Principios del análisis lineal predictivo

Aunque la codificación lineal predictiva se utilice ampliamente para extraer características de la señal de voz, esta se contempló originalmente como un problema independiente del reconocimiento y de la síntesis de la voz, se deseaba codificar la voz para poder transmitirla por un canal de comunicación sin demandar mucho ancho de banda, también el poder almacenar la señal de voz sin que requiera mucho espacio de disco.

La codificación lineal predictiva está basada en la idea de que una muestra de voz puede predecirse mediante una combinación lineal de las muestras anteriores de voz dentro de un cierto intervalo, para poder hacer eso, es

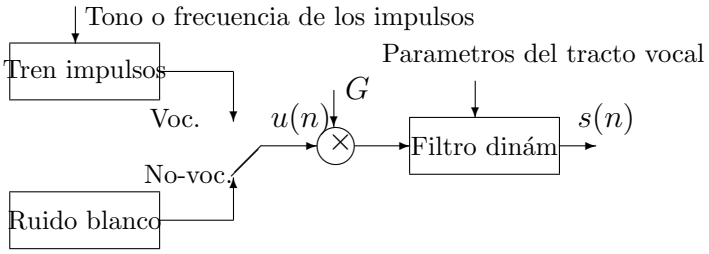

Figura 6.1: Modelo de producción de voz basado en la codificación lineal predictiva

necesario determinar los coeficientes predictores, esto se logra minimizando la suma de los cuadrados de las diferencias entre las muestras obtenidas mediante la predicción y las muestras reales.

La voz puede ser modelada como la salida de un sistema lineal cuyos parámetros varían en el tiempo y cuya entrada es un tren de impulsos o bien la salida de un generador de ruido dependiendo de si se va a producir un sonido vocalizado o un sonido no-vocalizado. Los parámetros del modelo de la Figura 6.1 son: la frecuencia F del tren de impulsos; la ganancia G; la posición del interruptor vocalizado/no-vocalizado y los coeficientes a_k, todos estos parámetros varían con el tiempo aunque no presentan cambios bruscos, excepto el interruptor por supuesto.

El tracto vocal es modelado por un filtro de puros polos cuya función de transferencia es:

$$\frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{6.1}$$

Aplicando transformada Z inversa:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{6.2}$$

Como se dijo antes, un predictor lineal predice una muestra de voz en base a las p muestras anteriores, así:

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) + Gu(n) \tag{6.3}$$

El error de predicción es la diferencia entre las muestras reales de voz y las muestras que produce el predictor:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) + Gu(n) \tag{6.4}$$

Al comparar (6.2) y (6.4) concluimos que para poder afirmar que los coeficientes predictores y los parámetros del tracto vocal coinciden, es decir $\alpha_k = a_k$ entonces deberá cumplirse $e(n) = Gu(n)$ lo cual significa que e(n) consiste de un tren de impulsos, o sea que $e(n)$ vale cero la mayor parte del tiempo. El error de predicción de tiempo corte se define como:

$$E_n = \sum_m e_n^2(m) = \sum_m [s_n(m) - \sum_{k=1}^{p} s_n(m-k)]^2 \tag{6.5}$$

donde $s_n(m)$ es la m-ésima muestra contada desde el inicio del marco que comienza en la n-ésima muestra, es decir la muestra $s(n+m)$.

Para encontrar los valores de α_k que minimizan E_n, hacemos $\partial E_n / \partial \alpha_i = 0$ para todo $i = 1, 2, ..., p$.

$$2 \sum_m \Big[S_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k) \Big] \frac{\partial \big[- \sum_{k=1}^{p} \alpha_k s_n(m-k) \big]}{\partial \alpha_i} = 0 \quad i = 1, 2, ..., p \tag{6.6}$$

pero:

$$\frac{\partial \big[- \sum_{k=1}^{p} \alpha_k s_n(m-k) \big]}{\partial \alpha_i} = -s_n(m-i) \quad i = 1, 2, ..., p \tag{6.7}$$

por lo cual:

$$2 \sum_m \Big[s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k) \Big] [-s_n(m-i)] = 0 \quad i = 1, 2, ..., p \tag{6.8}$$

de donde:

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^{p} \alpha_k \sum_m s_n(m-i)s_n(m-k) \;\; i = 1, 2, ..., p \quad (6.9)$$

Definiendo:

$$\phi_n(i,k) = \sum_m s_n(m-i)s_n(m-k) \quad (6.10)$$

Podemos escribir (6.9) en forma compacta como:

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi_n(i,0) \;\; i = 1, 2, ..., p \quad (6.11)$$

6.1.1. Método de la Autocorrelación

Para encontrar los coeficientes predictores que minimizan el error se deben calcular los coeficientes $\phi_n(i,k)$ para todo $1 \leq i \leq p$ y $0 \leq k \leq p$ y luego resolver el sistema de ecuaciones dado por (6.11). Para lograr esto, existen varios métodos, entre ellos el método de la covarianza, y el de la autocorrelación, describiremos aquí este último. En las ecuaciones anteriores, los límites de las sumatorias que utilizan a m como índice se han dejado sin especificar, sin embargo, al referirnos al error de predicción de tiempo corto, el intervalo coincide con el ancho N de los marcos, así, s(n+m)w(m) vale cero fuera del intervalo $[0, N-1]$ y el error de predicción de tiempo corto solo tiene valores distintos de cero dentro del intervalo $[0, N+p-1]$, por esta razón (6.11) se convierte en:

$$\phi_n(i,k) = \sum_{m=0}^{N+p-1} s_n(m-i)s_n(m-k) \quad i = 1, 2, ..., p \quad k = 0, 1, 2, ..., p \quad (6.12)$$

Haciendo $r = m - i$, tenemos que $m - k = r + i - k$

$$\phi_n(i,k) = \sum_{r=0}^{N-1-(i-k)} s_n(r)s_n(r+i-k) \quad i = 1, 2, ..., p \quad k = 0, 1, 2, ..., p \quad (6.13)$$

En la ecuación anterior, podemos apreciar que $\phi_n(i,k)$ es idéntica a la función de autocorrelación evaluada en $(i-k)$. En vista de que la función de autocorrelación es una función par, se cumple que:

$$\phi_n(i,k) = R_n(|i-k|) \tag{6.14}$$

Entonces (6.11) se puede expresar como:

$$\sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i) \;\; i = 1,2,...,p \tag{6.15}$$

6.2. Solución de las ecuaciones LPC

Para encontrar los coeficientes LPC, debemos resolver el sistema de ecuaciones expresado en forma compacta mediante (6.15), este sistema de ecuaciones podemos expresarlo en forma desarrollada:

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & ... & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & ... & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & ... & R_n(p-3) \\ : & : & : & ... & : \\ R_n(p) & R_n(p-1) & R_n(p-2) & ... & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ : \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ : \\ R_n(p) \end{bmatrix} \tag{6.16}$$

En forma compacta, este sistema de ecuaciones se puede representar como $T\alpha = r$

La matriz de autocorrelación T es una matriz de Topeplitz dado que es simétrica y los valores a lo largo de cualquier diagonal son un mismo valor. Los métodos mas conocidos para solucionar un sistemas de ecuaciones con estas características son los de Levinson y Robinson, pero al método más conocido se le conoce como *procedimiento recursivo de Durbin*

6.2.1. Método recursivo de Durbin para solucionar las ecuaciones de autocorrelación

Este procedimiento comienza con un predictor de primer orden, es decir, de un solo coeficiente e incrementa el orden recursívamente utilizando las

soluciones de orden inferior para obtener soluciones de órdenes superiores. La notación α_i^p denota al i-ésimo coeficiente de un predictor de orden p.

El algoritmo de Durbin resuelve primero un sistema predictor de primer orden ($p = 1$). El sistema de ecuaciones se reduce a:

$$\alpha_1^{(1)} R_n(0) = R_n(1) \tag{6.17}$$

de donde $\alpha_1^{(1)} = R_n(1)/R_n(0)$

El error de predicción para el predictor de primer orden se calcula mediante:

$$E_n = R_n(0) - \sum_{k=1}^{p} \alpha_k R_n(k) \tag{6.18}$$

el cual para un predictor de primer orden se reduce a:

$$E_n^{(1)} = R_n(0) - \alpha_1^{(1)} R_n(1) \tag{6.19}$$

Ahora se procede a solucionar un sistema predictor de segundo orden ($p = 2$), el sistema de ecuaciones a solucionar es:

$$\begin{bmatrix} R_n(0) & R_n(1) \\ R_n(1) & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1^{(2)} \\ \alpha_2^{(2)} \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \end{bmatrix} \tag{6.20}$$

Solucionando el sistema de ecuaciones para $\alpha_2^{(2)}$, tenemos:

$$\alpha_2^{(2)} = \frac{R_n(0)R_n(2) - R_n(1)^2}{R_n^2(0) - R_n^2(1)} \tag{6.21}$$

Dividiendo numerador y denominador por $R_n(0)$

$$\alpha_2^{(2)} = \frac{R_n(2) - R_n(1)^2/R_n(0)}{R_n(0) - R_n^2(1)/R_n(0)} \tag{6.22}$$

Sustituyendo ahora el resultado al que se llegó con el predictor de primer orden tenemos:

$$\alpha_2^{(2)} = \frac{R_n(2) - R_n(1)\alpha_1^{(1)}}{R_n(0) - R_n(1)\alpha_1^{(1)}} \tag{6.23}$$

Pero sabemos que $E_n^{(1)} = R_n(0) - \alpha_1^{(1)} R_n(1)$, por tanto:

$$\alpha_2^{(2)} = \frac{R_n(2) - R_n(1)\alpha_1^{(1)}}{E_n^{(1)}} \tag{6.24}$$

Observemos como un coeficiente de predicción de segundo orden $\alpha_2^{(2)}$ se puede calcular con pocas operaciones aritméticas si se conocen ya los coeficientes del predictor del orden inmediato inferior (en este caso $\alpha_1^{(1)}$). Algo similar se hace para $\alpha_1^{(2)}$. Una vez que se ha solucionado el sistema de segundo orden se procede a solucionar el sistema de tercer orden y así sucesivamente hasta llegar al orden deseado. El algoritmo de Levinson-Durbin se detalla a continuación:

1. $E^{(0)} = R_n(0)$
2. Desde $i = 1$ hasta p
   $k_i = \left[R_n(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R_n(i-j)\right] / E^{(i-1)}$
   $\alpha_i^{(i)} = k_i$
   Desde $j = 1$ hasta $i-1$ $\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}$
   Si $i < p$ entonces $E^{(i)} = (1 - k_i^2) E^{(i-1)}$

En el proceso de calcular los coeficientes predictores para un predictor de orden p se deben calcular los coeficientes de todos los predictores de órdenes inferiores a p. El error de predicción se puede estar monitoreando así como los coeficientes PARCOR (k_i).

Como el espectro de voz a ser analizado se puede representar adecuadamente con 2 polos (Ej. una pareja de polos complejos conjugados) por kilohertz, entonces se requiere un número de polos igual a la frecuencia de muestreo en kilohertz debido a la contribución del tracto vocal al espectro de la voz. Adicionalmente, se requieren 3 o 4 polos mas para representar adecuadamente la fuente de excitación (pulmones, valvula glotal y cuerdas vocales) y la radiación del sonido hacia el medio (labios). En vista de que el número de polos coincide con el grado del polinomio en el denominador de la función sistema, el número total de polos debe ser igual al p. Por ejemplo, para una frecuencia de muestreo de 8 KHz se recomienda $p = 12$. En la Figura 6.2, podemos observar como el rms del error de predicción disminuye al aumentar p, también el hecho de que el error de predicción siempre es mayor para sonidos no-vocalizados que para sonidos vocalizados.

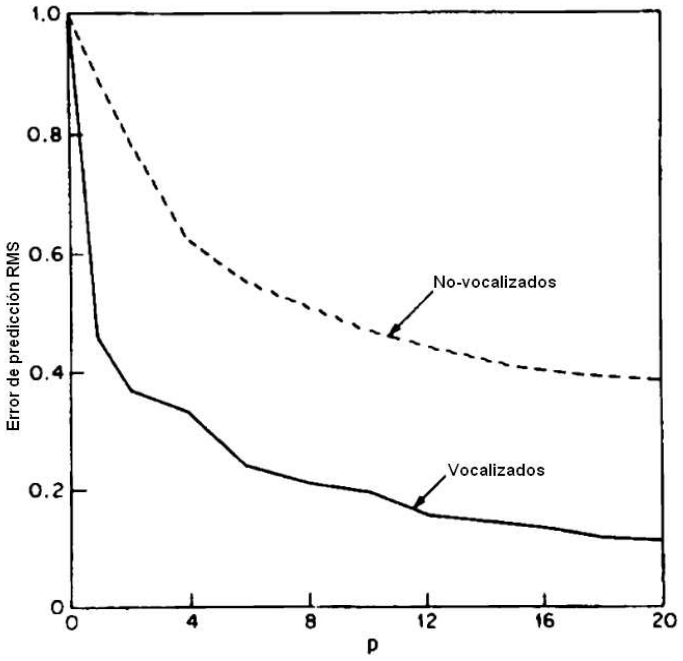

Figura 6.2: RMS del Error de predicción Vs p

6.3. La señal de error y su relación con la estimación del pulso glotal y aplicación a identificación de individuos por su voz

Como un producto colateral, el análisis LPC genera la señal de error $e(n)$ definida como:

$$e(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) = Gu(n) \tag{6.25}$$

Esto quiere decir que la señal de error es una buena aproximación de la fuente de excitación de un sistema de producción de voz el cual es modelado por un predictor de orden p. Basado en este razonamiento, es de esperarse que el error de predicción aumente drásticamente cada vez que inicia otro periodo del tono para sonidos vocalizados como podemos apreciar en la Figura 6.3. Por lo tanto dicho periodo (el tono) se puede determinar detectando las posiciones de las muestras de la señal $e(n)$ que son de mayor amplitud y midiendo entonces la diferencia de tiempo que hay entre muestras de $e(n)$ que superen un valor umbral. Alternativamente, se puede estimar el tono determinando la autocorrelación de la señal de error y detectando la posición del pico mas alto.

La razón por la que la señal de error es tan útil para la estimación del tono es porque el espectro de la señal de error es prácticamente plano ya que los formantes han sido eliminados y no aparecen en la señal de error.

El error de predicción medio cuadrático normalizado para el método de la autocorrelación se define como:

$$V_n = \frac{\sum_{m=0}^{N+p-1} e_n^2(m)}{\sum_{m=0}^{N-1} s_n^2(m)} \tag{6.26}$$

Definiendo $\alpha_0 = -1$, la señal de error de predicción (6.25) se puede expresar como:

$$e(n) = -\sum_{k=0}^{p} \alpha_k s(n-k) \tag{6.27}$$

Sustituyendo (6.27) en (6.26) y usando la definición de $\phi_n(i,j)$ tenemos:

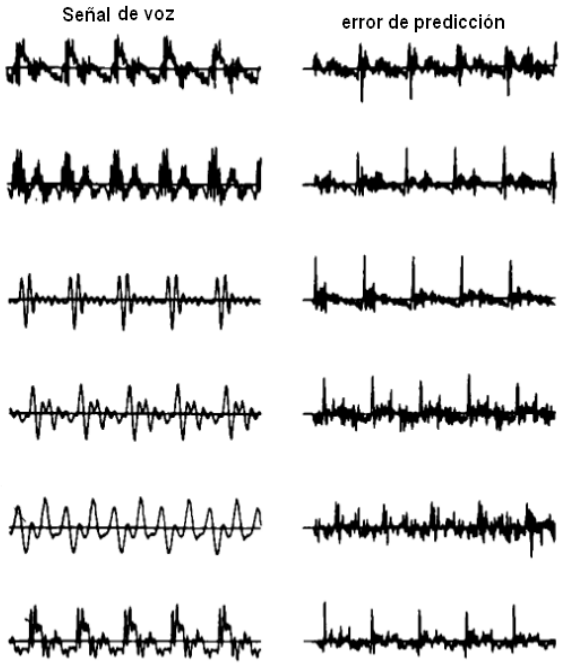

Figura 6.3: La señal de error

$$V_n = \sum_{i=0}^{p} \sum_{j=0}^{p} \alpha_i \frac{\phi_n(i,j)}{\phi_n(0,0)} \alpha_j \tag{6.28}$$

Una ventaja mas de utilizar el Algoritmo de Durbin es que el error de predicción medio cuadrático normalizado lo podemos calcular de manera muy simple mediante:

$$V_n = \prod_{i=1}^{p} (1 - k_i^2) \tag{6.29}$$

En el modelo de producción de voz, cuando el sonido es vocalizado, la excitación $u(n)$ es un tren de impulsos. Del análisis expuesto aquí concluimos que la señal de error de predicción es proporcional a la excitación ($e(n) = Gu(n)$) real, sabemos que la excitación en un tracto vocal humano consiste de un tren de pulsos glotales y que la forma del pulso glotal cambia de un individuo a otro y por esa razón es muy utilizado para la identificación de individuos por su voz. En conclusión la forma de la señal de error para un sonido vocalizado específico se podría utilizar para la identificación de el emisor de la señal de voz bajo análisis.

6.4. Interpretación en el dominio de la frecuencia del análisis lineal predictivo

Hasta ahora hemos discutido la predicción lineal en términos de ecuaciones de diferencias y funciones de correlación, es decir, en el dominio del tiempo. Sin embargo, hemos apuntado que los coeficientes del predictor lineal son asumidos iguales a los coeficientes del polinomio del denominador de la función sistema que modela los efectos combinados de respuesta del tracto vocal, forma del pulso glotal y radiación. Por tanto, dado un conjunto de coeficientes podemos encontrar la respuesta a la frecuencia del modelo de producción de voz simplemente evaluando $H(z)$ para $z = e^{j\omega}$,

$$H(e^{j\omega}) = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k e^{j\omega k}} = \frac{G}{A(e^{j\omega})} \tag{6.30}$$

Si graficamos $H(e^{j\omega})$ como función de la frecuencia observaremos picos en las frecuencias que corresponden a los formantes igual que en las representaciones espectrales que utilizan la Transformada de Fourier. En conclusion,

el análisis lineal predictivo puede verse como una técnica de estimación del espectro de tiempo corto.

6.5. Interpretación en el dominio de la frecuencia del error de predicción

Para determinar los coeficientes de un predictor lineal, minimizamos el error de predicción de tiempo corto. El error de predicción puede hacerse arbritariamente pequeño si incrementamos el orden del predictor lo suficiente. Es fácil mostrar que:

$$\lim_{p\to\infty} |H(e^{j\omega})|^2 = |X(e^{j\omega})|^2 \tag{6.31}$$

Para mostrar como el espectro determinado con un predictor lineal para valores elevados de p se parece mucho al espectro determinado con la Transformada discreta de Fourier, en la Figura 6.4 se muestra espectro $20log_{10}|X(e^{j\omega})|$ obtenido al realizar un análisis basado en la Transformada de Fourier, el análisis se realizó para un segmento de 20 ms de una señal de voz muestreada a 20 KHz y multiplicada por una ventana de Hamming, la señal de voz corresponde a la *a*. En la misma figura se muestra el espectro de $20log_{10}|H(e^{j\omega})|$ de la misma señal de voz para un predictor de orden $p = 28$.

El orden p del predictor puede efectivamente controlar el grado de suavidad del espectro de la señal como se muestra en la Figura 6.5 donde se muestra un segmento de una señal de voz muestreada a 6 KHz, su transformada de Fourier y el espectro obtenido mediante predicción lineal para diferentes valores de p.

6.6. Relación entre el análisis predictivo y los modelos de tubos sin pérdidas

Com vimos anteriormente, es posible modelar la producción de la voz mediante la concatenación de N tubos sin pérdidas como los mostrados en la Figura 6.9(a). Los coeficientes de reflexión r_k están relacionados a las áreas de los tubos que se conectan en la k-ésima unión de tubos sin pérdidas mediante:

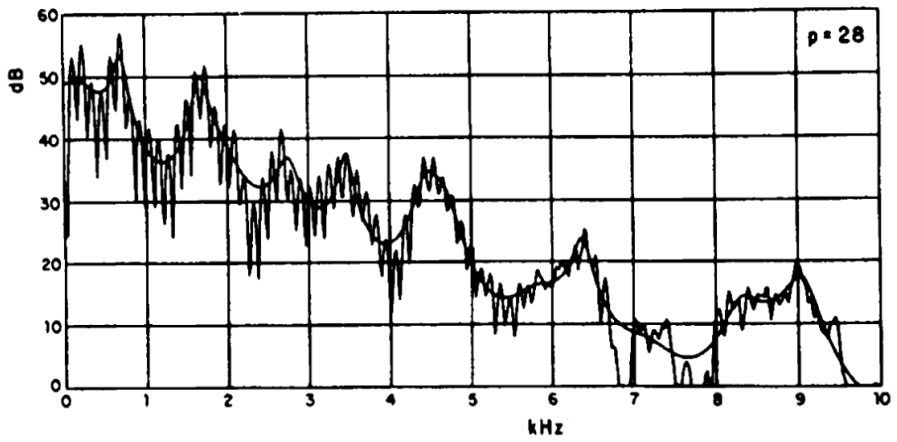

Figura 6.4: Espectro de la vocal *a* estimado con un predictor de orden $p = 28$ sobre el espectro de la misma señal obtenido por Transformada discreta de Fourier

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \tag{6.32}$$

La función de transferencia de tal sistema está condicionada a que el coeficiente de reflexión en el glotis sea de $r_G = 1$, asumiendo entonces una impedancia glotal infinita, se puede demostrar [3] que la función sistema del modelo reticular mostrado en la Figura 6.9 es:

$$V(z) = \frac{\prod_{k=1}^{N}(1 + r_k)z^{-N/2}}{D(z)} \tag{6.33}$$

donde $D(z)$ satisface la recursión:

$$\begin{aligned} D_0(z) &= 1 && (6.34)\\ D_k(z) &= D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1}) && (6.35)\\ D(z) &= D_N(z) && (6.36) \end{aligned}$$

Por otro lado, el polinomio

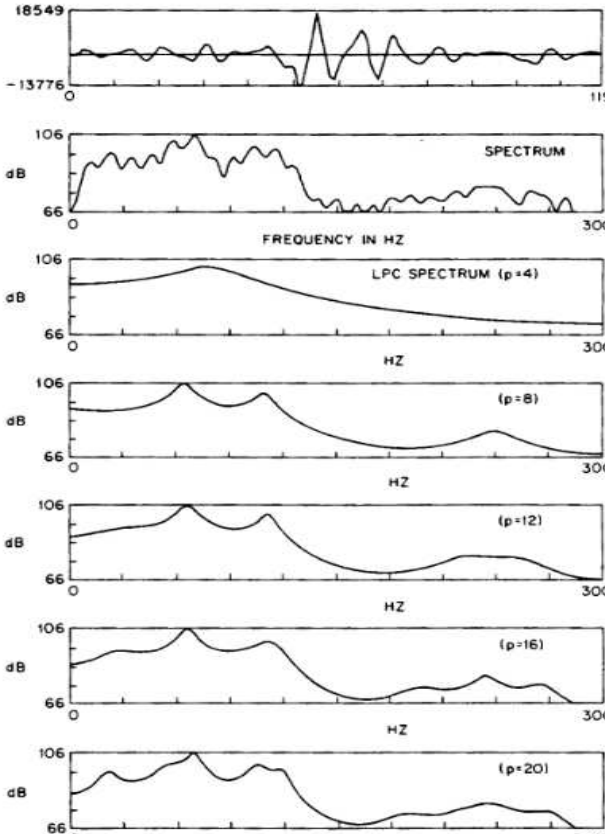

Figura 6.5: Espectro de la vocal a para diferentes valores del orden del predictor p

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{6.37}$$

que normalmente se obtiene mediante análisis de predicción lineal se podría obtener mediante la recursión:

$$\begin{aligned} A^0(z) &= 1 && (6.38)\\ A^i(z) &= A^{i-1}(z) - k_i z^{-i} A^{i-1}(z^{-1}) && (6.39)\\ A(z) &= A^p(z) && (6.40) \end{aligned}$$

donde los parámetros k_i son los llamados coeficientes PARCOR. Comparando las dos recursiones anteriores resulta claro que la función sistema

$$A(z) = \frac{G}{A(z)} \tag{6.41}$$

del modelo de producción de voz basado en predicción lineal tiene la misma forma que la función sistema de un modelo de tubos sin pérdidas que consiste de p secciones siempre y cuando:

$$r_i = -k_i \tag{6.42}$$

Es fácil mostrar también que las áreas de los tubos concatenados en la k-ésima unión se relacionan al coeficiente de reflexión correspondiente mediante:

$$A_{i+1} = A_i \frac{1 - k_i}{1 + k_i} \tag{6.43}$$

Observe que los coeficientes PARCOR proporcionan una razón entre áreas de secciones adyacentes, entonces el modelo equivalente de tubos no está determinado de manera absoluta de manera que habrá una infinidad de modelos de tubos concatenados con la misma función de transferencia.

Finalmente, si se realiza un pre-enfatizado de la señal previo al análisis de predicción lineal para remover los efectos debidos al pulso glotal y a la radiación, entonces las áreas resultantes son muy similares a aquellas que se pueden medir directamente de configuraciones del tracto vocal correspondientes.

6.7. Los coeficientes PARCOR

A los valores k_i que se determinan en de manera colateral por el Algortimo de Durbin-Levinson, se les conoce como *coeficientes PARCOR* debido a que son una medida de la correlación parcial (PARtial CORrelation) entre el error de predicción hacia adelante y el error de predicción hacia atrás, un concepto que utilizan lo métodos basados en retículas para determinar los coeficientes predictores.

Los coeficientes PARCOR son muy importantes puesto que el hecho de que se cumpla:

$$-1 < k_i < 1 \quad \forall 1 \leq i \leq p \tag{6.44}$$

es suficiente para estar seguros de que todas los polos de la función sistema del modelo del tracto vocal caen dentro del círculo trigonométrico unitario del plano complejo, esto implica un predictor estable. Tal garantía de estabilidad es una gran ventaja del método solución de autocorrelación que no existe en el método de solución de la covarianza.

Si por errores de precisión (redondeo) se detecta en algún momento que no se cumple $-1 < k_i < 1$, entonces se pueden revisar todos los polos y aquel que esté fuera del círculo trigonométrico unitario se puede reflejar hacia dentro del círculo sin menoscabo de la validez de los coeficientes pero garantizando la estabilidad del filtro.

Los coeficientes LPC pueden obtenerse a partir de los coeficientes PARCOR mediante la siguiente recursión:

$$a_i^{(i)} = k_{(i)} \tag{6.45}$$

$$a_j^{(i)} = k_i a_{i-j}^{(i-1)} \quad 1 \leq j \leq i-1 \tag{6.46}$$

estas ecuaciones se resuelven para $i = 1, 2, ..., p$ y los coeficientes del filtro son finalmente:

$$\alpha_j = a_j^{(p)} \quad 1 \leq j \leq p \tag{6.47}$$

De manera similar, los coeficientes PARCOR pueden obtenerse a partir de los coeficientes LPC usando la siguiente recursión hacia atrás:

$$k_{(i)} = a_i^{(i)} \tag{6.48}$$

$$a_j^{(i-1)} = \frac{a_j^{(i)} + a_i^{(i)} a_{i-j}^{(i)}}{1 - k_i^2} \quad 1 \leq j \leq i-1 \tag{6.49}$$

donde i va de p,$p-1$,.., a 1 estableciendo inicialmente:

$$a_j^{(p)} = \alpha_j \quad 1 \leq j \leq p \tag{6.50}$$

6.8. Síntesis de voz mediante parámetros LPC

La voz puede ser sintetizada a partir de los parámetros obtenidos en el análisis de predicción lineal de varias maneras. La manera mas simple tiene la misma representación paramétrica usada durante el análisis. La Figura 6.6 muestra un diagrama de bloques de tal sintetizador. Los parámetros de control variables en el tiempo que se requieren para el sintetizador son el tono, el booleano vocalizado/no-vocalizado, la ganacia o valor rms de la voz y los p coeficientes predictores. El generador de impulsos actúa como fuente de excitación para sonidos vocalizados y un generador de ruido blanco actúa como excitación para sonidos no-vocalizados produciendo muestras aleatorias no correlacionadas uniformemente distribuidas con media cero y desviación standard unitaria. Las muestras sintetizadas de voz se determinan mediante:

$$\tilde{s}(n) = \sum_{k=1}^{p} \tilde{s}(n-k) + Gu(n) \tag{6.51}$$

Para sonidos no-vocalizados los parámetros simplemente se cambian en cada marco Para sonidos vocalizados, los parámetros son cambiados al inicio de cada periodo, a esto se le conoce como síntesis sincronizada al tono y se ha descubierto que es una síntesis mas efectiva que cuando se cambian los parámetros en cada marco a lo cual se conoce como síntesis asíncrona. La síntesis síncrona requiere interpolación de los parámetros para descubrir que valor deben tener al inicio del periodo. El tono y la ganancia conviene interpolarlos geométricamente (linealmente en escala logarítmica), sin embargo, debido a restricciones de estabilidad, los parámetros de estabilidad no pueden interpolarse directamente. La razón es que la interpolación de dos conjuntos de coeficientes correspondientes a dos predictores estables puede

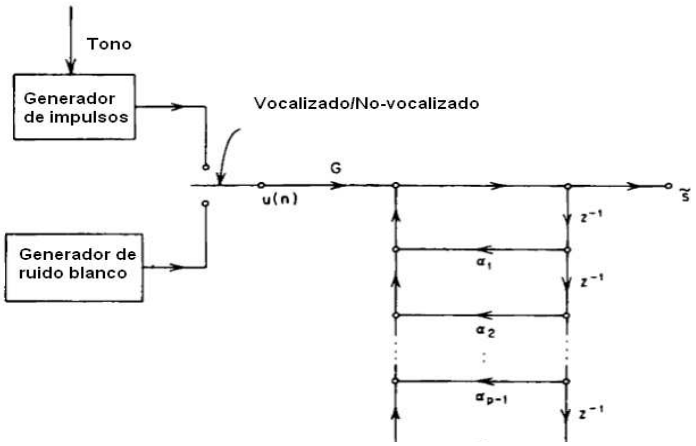

Figura 6.6: Diagrama de bloques de un sintetizador lineal predictivo

producir un conjunto de coeficientes de un predictor inestable. Una solución a este problema consiste en interpolar los primeros p valores de la función de autocorrelación de la salida sintetizada usando cada conjunto de coeficientes a interpolar. En lugar de interpolar directamente los coeficientes predictores, interpolamos los dos conjuntos de valores de autocorrelación obtenidos, una vez obtenido un conjunto de valores de autocorrelación producto de la interpolación, se procede a resolver el sistema mediante el método de Durbin garantizando así que el predictor producido será estable.

La principal ventaja del sintetizador mostrado en la Figura 6.6 es su sencillez, su principal desventaja es que requiere de gran precisión debido a su estructura recursiva directa que lo hace demasiado sensible a cambios de los coeficientes.

6.9. Determinación del tono usando los coeficientes LPC

Sabemos que la señal de error $e(n)$ obtenida en el análisis LPC puede ser usada para estimar el tono diréctamente. Aunque este método es capaz de

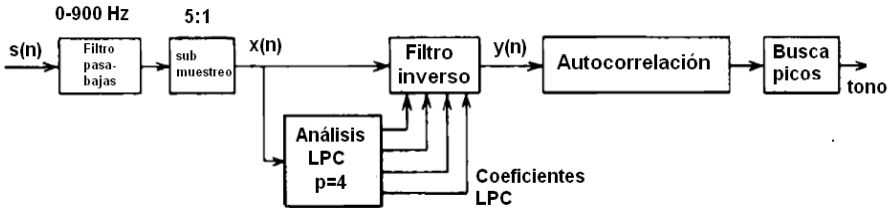

Figura 6.7: Diagrama de bloques de un detector de tono basado en análisis LPC

de encontrar el periodo correctamente, existe un método más sofisticado de detección del tono, se conoce como Algoritmo SIFT (Simple Inverse Filtering Tracking).

La Figura 6.7 muestra un diagrama de bloques del Algoritmo SIFT, la señal de entrada $s(n)$ es pasada por un filtro pasabajas con frecuencia de corte de 900 Hz, enseguida es submuestreada para eliminar la redundancia, si la frecuencia de muestreo fué de 10 KHZ, después del submuestreo tendremos una frecuencia de muestreo de solo 2 KHz, esto se consigue conservando solo una de cada cinco muestras y eliminando al resto. A la señal resultante $x(n)$ se le hace un análisis LPC mediante el método de autocorrelación para un predictor de orden $p = 4$ ya que un predictor de cuarto orden es suficiente para analizar una señal con un contenido de frecuencia de menos de 1 KHz. La señal es sintetizada a partir de los 4 coeficientes obtenidos en el paso anterior obteniendo $y(n)$, la cual es una señal con un espectro aproximadamente plano. Así pues, el propósito del análisis LPC es aplanar el espectro, enseguida se obtiene la función de autocorrelación de $y(n)$ ´de donde se localiza el pico mas grande, cuya ubicación es la estimación del tono.

En la Figura 6.8 se muestran algunas señales típicas obtenidas en las diferentes fases del análisis. En la Figura 6.8(a) se muestra la forma de onda analizada y la Figura 6.8(b) muestra su espectro en donde se aprecia que hay solo un formante en el rango de 250 Hz. La Figura 6.8(c) muestra el espectro de la señal que se obtiene a la salida del filtro inverso donde podemos ver que el espectro es aproximadamente plano, la Figura 6.8(d) muestra la señal

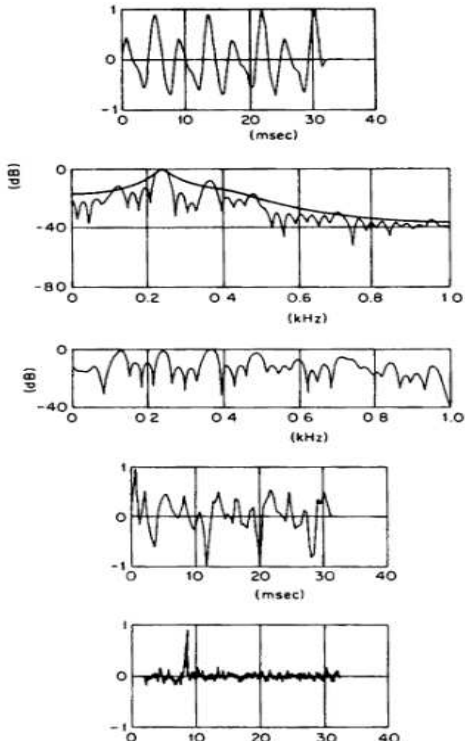

Figura 6.8: Señales involucradas en el Algoritmo SIFT para estimación del tono

sintetizada en el dominio del tiempo. Finalmente, la Figura 6.8(e) muestra la función de autocorrelación de la señal sintetizada de donde un tono de 8 ms aproximadamente se puede ver con claridad.

6.10. Análisis de formantes usando coeficientes LPC

Los formantes se pueden estimar de dos maneras a partir de los coeficientes LPC, la primera de ellas consiste en factorizar el polinomio predictor y una vez determinadas las raíces tratar de descubrir cuales son formantes y cuales son polos que le dan forma al espectro. Alternativamente, los formantes se pueden determinar buscando los primeros picos del espectro estimado a partir de los coeficientes LPC.

La ventaja de utilizar los coeficientes LPC para estimar los formantes es que es menos complicado encontrarlos dado que el espectro obtenido a partir de los coeficientes LPC es un espectro suavizado, la desventaja inherente a la estimación de los formantes a partir de los coeficientes LPC es que estos son obtenidos asumiendo que el tracto vocal puede modelarse mediante un filtro de puros polos y aunque esto es adecuado para efectos de clasificación, se sabe que el espectro de sonidos nasales no solo tiene polos sino ceros por lo que la estimación de los formantes a partir de coeficientes LPC no sería muy preciso para este tipo de sonidos.

6.11. Reconocedor de palabras aisladas basado en coeficientes LPC

El reconocedor de palabras aisladas en base a coeficientes LPC es prácticamente igual al sistema de reconocimiento de palabras aisladas mediante espectrogramas descrito en la sección 4.8. Ambos sistemas caracterizan una palabra mediante una matriz de flotantes, en ambos casos la matriz tiene un número fijo de columnas y un número variable de renglones pues este depende de la duración de la elocución correspondiente. La única diferencia entre ambos sistemas es el denominado *front end*, es decir , el módulo de extracción de características de la señal de voz y por supuesto la interpretación que debe de hacerse de dicha matriz, en el caso del reconocedor mediante espec-

trogramas descrito en 4.8 cada columna corresponde con una banda crítica de Bark, mientras que en el reconocedor hecho en base a coeficientes LPC, cada columna corresponde con uno de los parámetros del filtro dinámico de puros polos que modela al tracto vocal, es decir, con uno de los coeficientes predictores. Los demás módulos no es entonces necesario cambiarlos, para comparar matrices se puede usar el doblado dinámico en tiempo, se pueda usar el mismo criterio del vecino mas cercano o de k-vecinos y usar distancia Euclidiana o distancia coseno, sin embargo, conviene aprovechar que se trata de coeficientes LPC y utilizar la distancia de Itakura, la cual fué diseñada específicamente para comparar vectores cuyos componentes son precisamente coeficientes LPC, aunque esta distancia requiere que conservemos los valores de autocorrelación que se utilizaron para determinar los coeficientes LPC mediante el algoritmo de Durbin-Levinson, en realidad solo se requieren los valores de autocorrelación de donde se determinaron los coeficientes de uno de los dos vectores, entonces para un sistema de reconocimiento podemos usar los valores de autocorrelación correspondientes a los coeficientes LPC de la consulta, de esa manera no es necesario aumentar el espacio requerido para almacenar el diccionario. En cuanto a los parámetros del sistema, ya hemos mencionado que el orden del predictor debe de ser igual a la frecuencia de muestreo en KHz y tres o cuatro polos adicionales, por lo tanto si la frecuencia de muestreo que se usa es de 8KHz, se recomienda usar $p = 12$. Finalmente, un reconocedor hecho en base a coeficientes LPC no requiere de aplicación de ventanas (aunque tampoco le perjudica) y como no se aplicará la transformada rápida de Fourier no es necesario usar marcos cuyo tamaño en muestras sea una potencia de dos (ni rellenar con ceros), de todos modos conviene usar marcos de 30 ms con traslape de dos tercios de manera que los marcos avancen 10 ms cada vez.

6.11.1. Distancia de Itakura

Hemos visto como tanto para reconocimiento de voz como para identificación de individuos por su voz es necesario comparar cuantitativa y efeciéntemente dos marcos de voz para los cuales el análisis LPC ha extraído diferentes vectores de coeficientes LPC. Entonces, requerimos una medida de distancia entre marcos de voz $D(a, \hat{a})$ donde $a = (1, \alpha_1, \alpha_2, ..., \alpha_p)$ es el vector con los coeficientes LPC de un marco de voz y $\hat{a} = (1, \hat{\alpha_1}, \hat{\alpha_2}, ..., \hat{\alpha_p})$. Como $D(a, \hat{a})$ es una medida de distancia se requiere que

$$D(a, \hat{a}) \geq 0 \tag{6.52}$$

y que:

$$D(a, a) = 0 \tag{6.53}$$

Itakura propuso una distancia basado en el siguiente razonamiento: Debido al ruido y a la imprecisión del modelo de voz basado en predicción lineal, no es posible determinar los verdaderos valores LPC del segmento de voz en cuestión, solo estamos haciendo una estimación de los mismos. Suponga que se tiene un segmento de voz y los coeficientes estimados son los coeficientes $\hat{a}$, entonces el problema es determinar la probabilidad de que $\hat{a}$ haya sido extraído de un segmento de voz cuyos coeficientes LPC verdaderos son los de a.

La distribución de probabilidad de $\hat{a}$ es una gaussiana multivariada con media a y matriz de covarianzas Σ la cual se determina mediante:

$$\Sigma = \frac{R^{-1}}{N} \hat{a} R \hat{a}^t \tag{6.54}$$

donde R es pa matriz de correlación (de $p+1$ por $p+1$) del segmento de voz, N es la longitud del marco de voz en muestras.

Entonces, la probabilidad de obtener $\hat{a}$ es:

$$P(\hat{a}|a) = \frac{1}{(2\pi)^{p/2}|\Sigma^{1/2}} e^{-0{,}5(\hat{a}-a)\Sigma^{-1}(\hat{a}-a)^t} \tag{6.55}$$

En vista de que el logaritmo es una función monotónicamente creciente, podemos aplicarlo a la ecuación anterior y usar el resultado como medida de distancia luego de eliminar los términos constantes y expresar σ en términos de R y N

$$D(\hat{a}, a) = (\hat{a} - a)\Big[N \frac{R}{\hat{a} R a^t}\Big](\hat{a} - a)^t \tag{6.56}$$

Por supuesto una probabilidad pequeña implica una distancia grande y vicebersa, eso y algunas consideraciones respecto al ahorro de cómputos condujo a Itakura a proponer la siguiente medida de distancia.

$$D'(\hat{a}, a) = log\Big[\frac{a R a^t}{\hat{a} R \hat{a}^t}\Big] \tag{6.57}$$

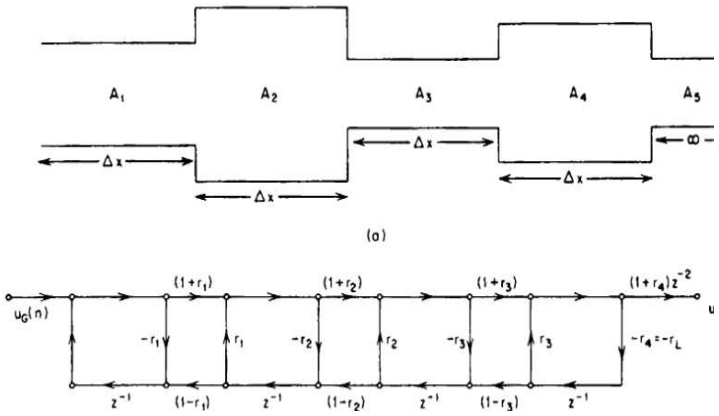

Figura 6.9: (a) Modelo de tubos sin pérdidas. (b) Grafo de flujo de señal correspondiente

La clave para entender esta distancia es que dos vectores de coeficientes LPC deben tener una distancia corta entre ellos si fueron extraídos de segmentos de voz que suenan muy parecido

6.12. Síntesis de voz basado en coeficientes LPC

La alternativa a la síntesis basada en coeficientes predictores que resulta mas atractiva es la síntesis basada en coeficientes de reflección del modelo basado en concatenación de tubos sin pérdidas o coeficientes PARCOR, en otras palabras se puede cambiar del sintetizador mostrado en la Figura 6.6 la parte que recibe $u(n)$ y entrega $\tilde{s}(n)$ por la red mostrada en la Figura 6.9(b). La ventaja principal es que los multiplicadores son los coeficientes de reflección $r_i = -k_i$, los cuales tienen la propiedad de estar acotados ($|k_i| < 1$) y también que pueden ser interpolados directamente produciendo filtros estables, esta estructura es también menos sensible a efectos de cuantización por implementación en tipos de datos de precisión finita.

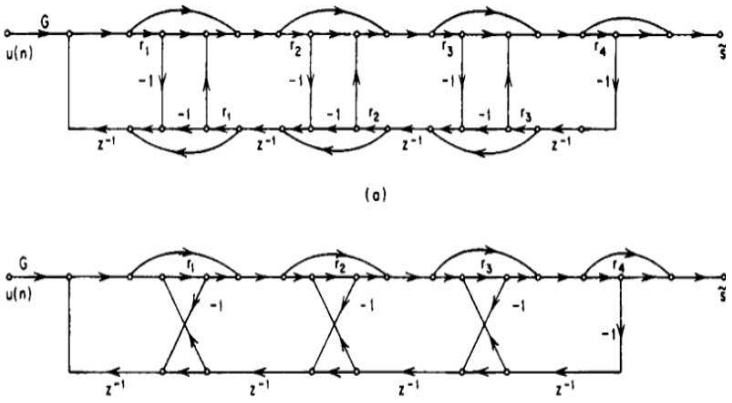

Figura 6.10: Modelos de tubo sin pérdidas equivalente usando (a) dos uniones multiplicadoras; y (b) una unión multiplicadora

Implementar el sintetizador de orden p de la Figura 6.9 requiere $4p + 2$ multiplicaciones y $2(p + 1)$ adiciones por muestra lo cual es una desventaja respecto al sintetizador del mismo orden de la Figura 6.6 que solo requiere p multiplicaciones y p adiciones, sin embargo, el sintetizador de la Figura 6.9(b) puede ser reemplazado por el de la Figura 6.10(a) que requiere $2p - 1$ multiplicaciones y $4p - 1$ adiciones o el de la Figura 6.10(b) que solo requiere p multiplicaciones y $3p - 2$ adiciones.

Bibliografía

[1] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing. Second Edition*, 1992.

[2] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing. Principles, Algorithms and Applications*. MacMillan, 1992.

[3] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall.

[4] C. Shannon and W. Weaver, *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley and Sons Inc., 2001.

[6] R. G. Campos and L. Z. Juarez, "A discretization of the continuos fourier transform," *Nuovo Cimento*, vol. 107, pp. 703–711, 1992.

[7] J. H. Mathews and K. D. Fink, *Numerical Methods with Matlab 3rd Edition*. Prentice Hall, 2000.

[8] R. G. Gonzalez, "A quadrature formula for the hankel transform," *Numerical Algorithms*, vol. 9, no. 3, pp. 343–354, 1995.

[9] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schioler, "Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music," in *International Symposium on Music Information Retrieval (ISMIR)*, 2006.