



# Síntesis y Reconocimiento de Voz

## *Facultad de Ingeniería Eléctrica*

Dr. José Antonio Camarena Ibarrola

camarena@umich.mx

Universidad Michoacana de San Nicolás de Hidalgo

Facultad de Ingeniería Eléctrica

# Fundamentos de procesamiento digital de señales. Introducción

El procesamiento de señales se puede definir como la transformación de una señal a una forma mas deseable.

$$y(n) = T[x(n)] \quad (1)$$

La clase especial de sistemas lineales e invariantes en el tiempo (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

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) = x(n) * h(n) \quad (2)$$

donde  $h(n)$  es la respuesta del sistema al impulso unitario. En forma equivalente:

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) = h(n) * x(n) \quad (3)$$

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

# La Transformada Z. Definición y propiedades

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

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \quad (4)$$

La transformada Z inversa se define como:

$$x(n) = \frac{1}{2\pi j} \oint_C X(z)z^{n-1} dz \quad (5)$$

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

Condición de convergencia:

$$\sum_{n=-\infty}^{\infty} |x(n)||z^{-n}| < \infty \quad (6)$$

El conjunto de valores  $z$  donde la serie converge se denomina “Región de Convergencia” (ROC) y normalmente tiene la forma:

$$R_1 < |z| < R_2$$

# La Transformada Z. Definición y propiedades

Ejemplo 1  $x(n) = \delta(n - n_0)$

$$X(z) = \sum_{n=-\infty}^{\infty} \delta(n) z^{-n} = z^{-n_0} \quad (8)$$

# La Transformada Z. Definición y propiedades

Ejemplo 2  $x(n) = u(n) - u(n - N)$

$$X(z) = \sum_{n=-\infty}^{\infty} [u(n) - u(n - N)]z^{-n} = \sum_{n=0}^{N-1} z^{-n} \quad (9)$$

$$X(z) = 1 + \frac{1}{z} + \dots + \frac{1}{z^{N-1}} \quad (10)$$

Dado que:

$$\sum_{n=0}^{N-1} r^n = \begin{cases} \frac{1-r^N}{1-r} & r \neq 1 \\ N & r = 1 \end{cases} \quad (11)$$

entonces:

$$X(z) = \frac{1 - z^{-N}}{1 - z^{-1}} \quad (12)$$

# La Transformada Z. Definición y propiedades

Ejemplo 3  $x(n) = a^n u(n)$

$$X(z) = \sum_{n=-\infty}^{\infty} a^n u(n) z^{-n} = \sum_{n=0}^{\infty} a^n z^{-n} \quad (13)$$

$$X(z) = \sum_{n=0}^{\infty} \left(\frac{a}{z}\right)^n = \lim_{N \rightarrow \infty} \frac{1 - (a/z)^N}{1 - a/z} \quad (14)$$

Este límite existirá si se cumple  $|a| < |z|$  (Región de Convergencia)

$$\lim_{N \rightarrow \infty} \left(\frac{a}{z}\right)^N = 0 \quad (15)$$

Por lo tanto:

$$X(z) = \frac{1}{1 - a/z} \quad (16)$$

# La Transformada Z. Definición y propiedades

Ejemplo 4  $x(n) = b^n u(-n - 1)$

$$X(z) = \sum_{n=-\infty}^{\infty} b^n u(-n - 1) z^{-n} = \sum_{n=-\infty}^{-1} b^n z^{-n} \quad (17)$$

$$X(z) = \sum_{n=1}^{\infty} b^{-n} z^n = \sum_{n=1}^{\infty} \left(\frac{z}{b}\right)^n = X(z) = \sum_{n=0}^{\infty} \left(\frac{z}{b}\right)^n - 1 \quad (18)$$

$$X(z) = \lim_{N \rightarrow \infty} \frac{1 - (z/b)^N}{1 - z/b} - 1 \quad (19)$$

Este límite existirá si se cumple  $|z| < |b|$ , (Región de convergencia)

$$\lim_{N \rightarrow \infty} \left(\frac{z}{b}\right)^N = 0 \quad (20)$$

$$X(z) = \frac{1}{1 - z/b} - 1 \quad (21)$$

## La Transformada Z. Definición y propiedades

### Región de Convergencia

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  $|R_1| < |z| < |R_2|$



# La Transformada Z. Definición y propiedades

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

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	$x(n)w(n)$	$\frac{1}{2\pi j} \oint_C X(v)W(z/v)v^{-1} dv$

# La Transformada Z. Definición y propiedades

Demostración de la Propiedad de desplazamiento: Sea

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \quad (22)$$

$$\sum_{n=-\infty}^{\infty} x(n - n_0)z^{-n} = \sum_{m=-\infty}^{\infty} x(m)z^{-m-n_0} = z^{-n_0} \sum_{m=-\infty}^{\infty} x(m)z^{-m} = z^{-n_0} X(z) \quad (23)$$

donde  $m = n - n_0$

# La Transformada de Fourier. 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:

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \quad (24)$$

$$x(n) = \int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n} d\omega \quad (25)$$

Observe que la integral de contorno se convirtió en una integral de  $-\pi$  a  $\pi$  que realmente está integrando los valores alrededor del círculo trigonométrico unitario. Basta sustituir  $e^{j\omega}$  por  $z$  en la Tabla para obtener un conjunto de propiedades de la Transformada de Fourier.

# 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:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N} \quad (26)$$

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

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)e^{j2\pi kn/N} \quad (27)$$

# La Transformada Discreta de Fourier

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

$$\begin{bmatrix} X(0) \\ X(1) \\ X(2) \\ \vdots \\ X(N) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & W_N^1 & W_N^2 & \dots & W_N^{N-1} \\ 1 & W_N^2 & W_N^4 & \dots & W_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & W_N^{N-1} & W_N^{2(N-1)} & \dots & W_N^{(N-1)^2} \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ \vdots \\ x(N) \end{bmatrix}$$

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



# La Transformada Discreta de Fourier

- La complejidad para determinar directamente la DFT es  $O(N^2)$ .
- La transformada discreta de Fourier puede de hecho determinarse con  $O(N \log_2 N)$  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.
- Sabemos de varios métodos de cálculo eficiente de la DFT, el de Danielson and Lanczos in 1942, proveé una de las mas claras derivaciones del algoritmo.

# La Transformada Discreta de Fourier

El lema de Danielson y Lanczos

La DFT de una secuencia de longitud  $N$  puede expresarse como la suma de dos DFTs cada una de longitud  $N/2$ , conformadas por los valores de las posiciones pares e impares respectivamente.

$$X(k) = \sum_{n=0}^{N-1} e^{-j2\pi kn/N} x(n) \quad (28)$$

$$X(k) = \sum_{n=0}^{N/2-1} e^{-j2\pi k(2n)/N} x(2n) + \sum_{n=0}^{N/2-1} e^{-j2\pi k(2n+1)/N} x(2n+1) \quad (29)$$

$$X(k) = \sum_{n=0}^{N/2-1} e^{-j2\pi kn/(N/2)} x(2n) + W_N^k \sum_{n=0}^{N/2-1} e^{-j2\pi kn/(N/2)} x(2n+1) \quad (30)$$

$$X(k) = X^e(k) + W_N^k X^o(k) \quad (31)$$

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

# La Transformada Discreta de Fourier

- El problema de calcular la DFT de una señal de longitud  $N$  se convierte en el problema de calcular dos DFTs  $X^e(k)$  y  $X^o(k)$  cada una de longitud  $N/2$ .
- El problema de determinar  $X^e(k)$  y  $X^o(k)$  se convierte en el problema de calcular 4 DFTs, cada una de longitud  $N/4$ .
- Denotamos  $X^{ee}(k)$  y  $X^{eo}(k)$  como las DFTs de los valores ubicados en posiciones par-par y par-impar en las sucesivas subdivisiones de la señal.
- $N$  debe ser 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.
- Seguir aplicando el lema de Danielson-Lanczos hasta que la transformada sea de longitud 1.
- La DFT de longitud uno es simplemente la operación identidad.
- 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.

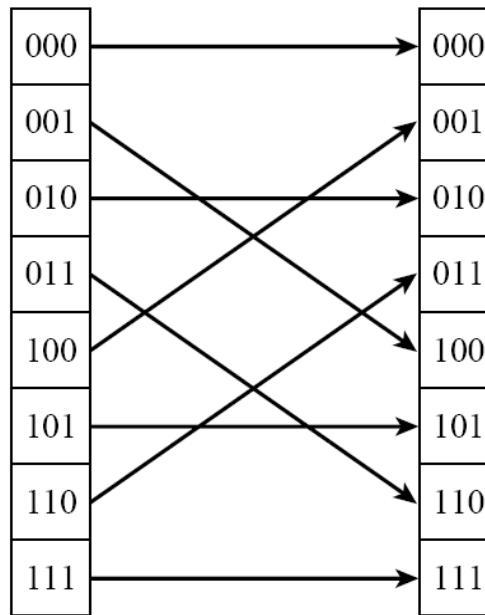
$$X^{eoeoeoeo\dots oee}(k) = x(n) \quad (32)$$

para alguna  $n$

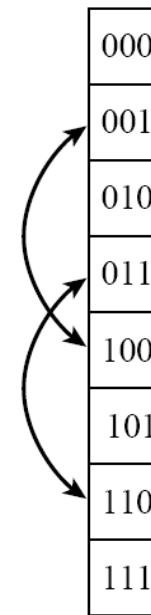


# La Transformada Discreta de Fourier

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  $n$ .



(a)



(b)

# La Transformada Discreta de Fourier

La idea de la inversión de bits junto con el lema de Danielson-Lanczos sirve para hacer un algoritmo práctico de FFT:

- Tomamos el vector original de datos  $x$  y lo reorganizamos en el orden de bits invertidos.
- 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,
- combinamos parejas de parejas para obtener transformadas de “señales” de longitud 4 y así sucesivamente hasta que la primera mitad del arreglo se combine con la segunda mitad del arreglo para obtener la transformada final.
- Cada combinación toma del orden de  $N$  operaciones y como hay  $\log_2 N$  combinaciones, entonces el algoritmo completo es del orden  $N \log_2 N$ .
- En la implementación de este algoritmo se debe cuidar de no llamar a la función seno o a la coseno más 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.



# La Transformada Discreta de Fourier.

## Propiedades y Limitaciones

---

- Si se aplica la Transformada Discreta de Fourier requiere a señales que no son estacionarias se pueden obtener resultados inesperados
- Decimos que una señal no es estacionaria si los contenidos de frecuencia de la señal cambian en el tiempo.
- 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

# La Transformada Discreta de Fourier.

## Propiedades y Limitaciones

Ejemplo, Sea

$$x(t) = \sin\left(\frac{t^2}{2}\right) \quad (33)$$

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.

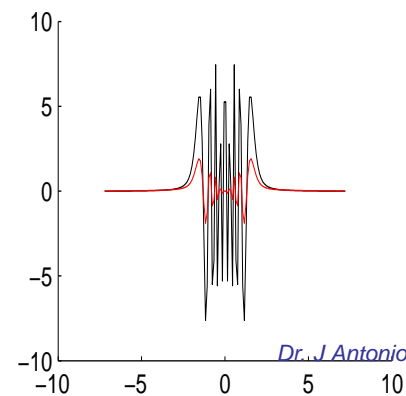
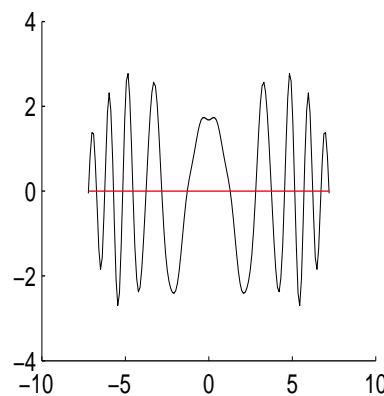
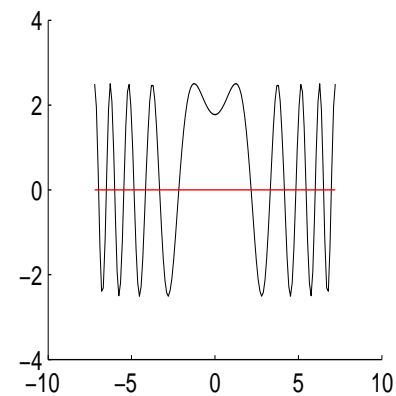
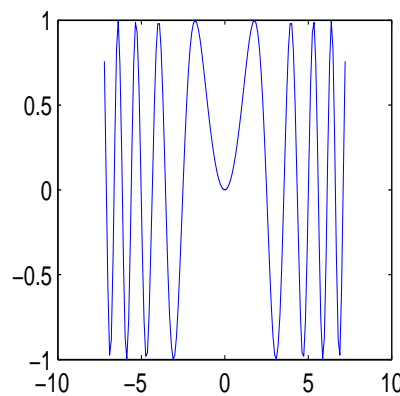
$$X(j\omega) = \int_{-\infty}^{\infty} \sin\left(\frac{t^2}{2}\right) e^{-j\omega t} dt \quad (34)$$

$$= \pi \left[ \cos\left(\frac{\omega^2}{2}\right) + \sin\left(\frac{\omega^2}{2}\right) \right] \quad (35)$$

# La Transformada Discreta de Fourier.

## Propiedades y Limitaciones

La señal no estacionaria  $x(t) = \sin(\frac{t^2}{2})$ , su CFT, una Discretización de la CFT y la DFT, con rojo la parte imaginaria y en negro la parte real, la parte imaginaria de la CFT es cero. La DFT falla al entregar una parte imaginaria diferente de cero.



# 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:

$$f(Hz) = \frac{(k)(f_m)}{N} \quad (36)$$

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

# Diseño de un filtro usando la DFT y la DFT inversa

- 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
- 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.

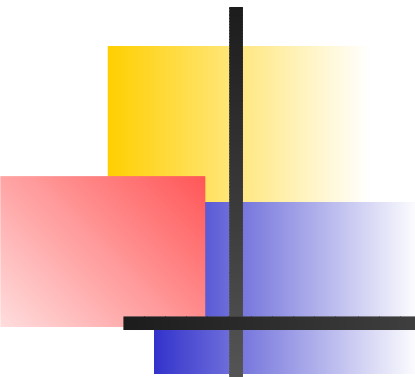
# Filtros Digitales. Causalidad, Linealidad, Sistemas invariantes en el tiempo

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:

$$Y(z) = H(z)X(z) \quad (37)$$

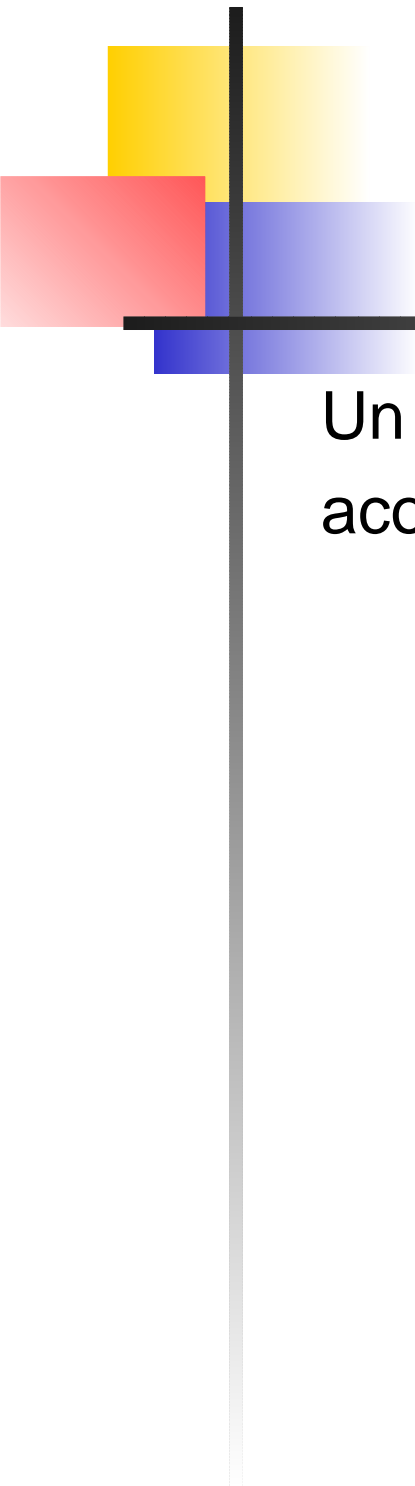
$H(z)$ , la llamada función sistema es la transformada Z de la respuesta del sistema al impulso unitario  $h(n)$ , mientras que a  $H(e^{j\omega})$  (la transformada de Fourier de  $h(n)$ ) se le conoce como su respuesta a la frecuencia.





## Filtros Digitales. Causalidad, Linealidad, Sistemas invariantes en el tiempo

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  $n < 0$ .



# Filtros Digitales. Causalidad, Linealidad, Sistemas invariantes en el tiempo

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

# Filtros Digitales. Causalidad, Linealidad, Sistemas 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:

$$y(n) - \sum_{k=1}^N a_k y(n-k) = \sum_{r=0}^M b_r x(n-r) \quad (38)$$

Aplicando transformada Z a ambos lados obtenemos:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{r=0}^M b_r z^{-r}}{1 - \sum_{k=1}^N a_k z^{-k}} \quad (39)$$

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:

$$H(z) = \frac{A \prod_{r=0}^M (1 - c_r z^{-1})}{\prod_{k=1}^N (1 - d_k z^{-1})} \quad (40)$$

# Filtros FIR

Si todos los coeficientes  $a_k$  valen cero la ecuación de diferencias se convierte en:

$$y(n) = \sum_{r=0}^M b_r x(n-r) \quad (41)$$

La respuesta al impulso es :

$$h(n) = \begin{cases} b_n & 0 \leq n \leq M \\ 0 & \text{de otro modo} \end{cases} \quad (42)$$

$H(z)$  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:

$$h(n) = \pm h(M-n) \quad (43)$$

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

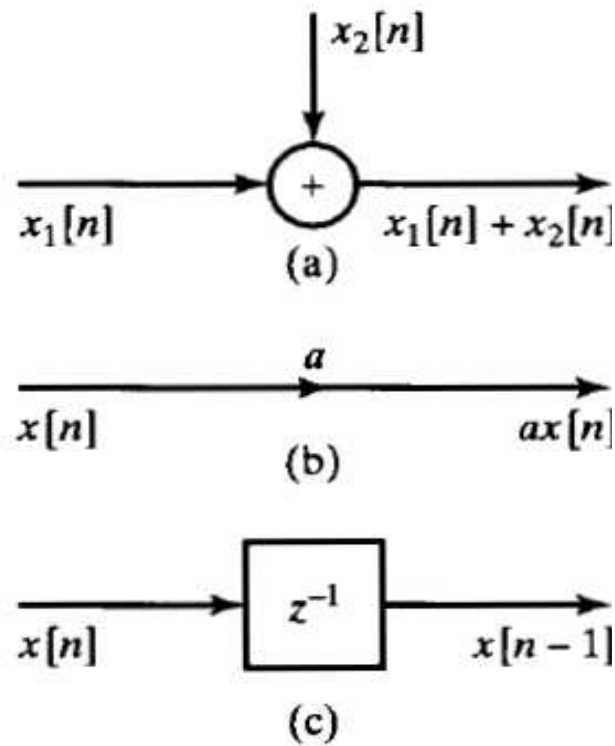


## Filtros FIR

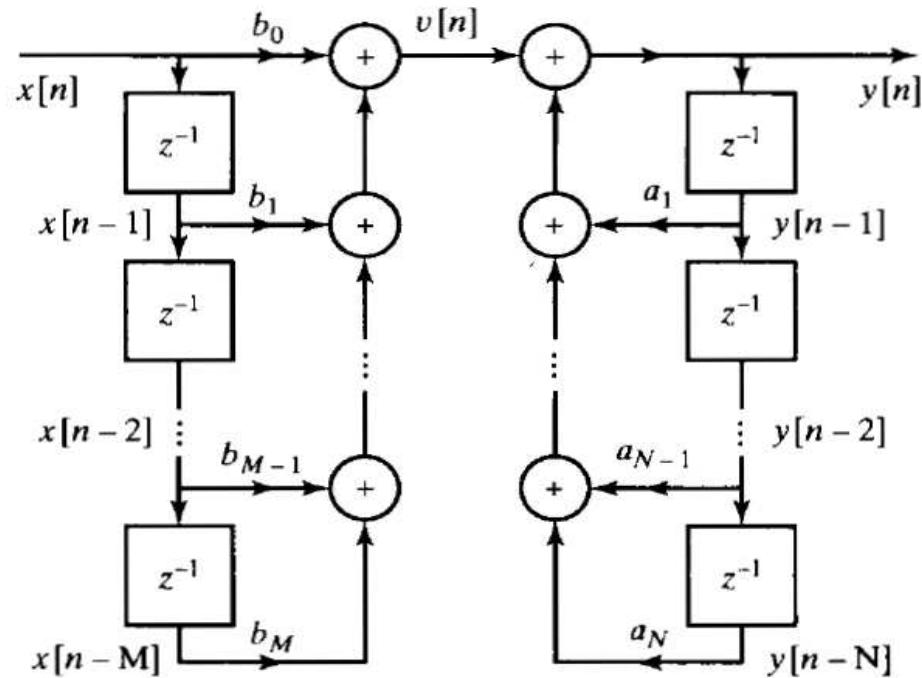
---

- 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.

# Filtros FIR



# Ecuación de diferencias como diagrama de bloques





## Filtros FIR

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

$$H(\omega) = \sum_{k=0}^{M-1} b_k e^{-j\omega k} \quad (44)$$

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



# Diseño de un filtro FIR. Determinación de los coeficientes mediante la técnica de 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:

$$h(n) = h(M - 1 - n) \quad (45)$$

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

$$\omega_k = \frac{2\pi k}{M} \quad \forall k = \begin{cases} 0, 1, \dots, \frac{M-1}{2} & M \text{ impar} \\ 0, 1, \dots, \frac{M}{2} - 1 & M \text{ par} \end{cases} \quad (46)$$

# Diseño de un filtro FIR. Determinación de los coeficientes mediante la técnica de muestreo de la frecuencia

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

$$\sum_{n=0}^{(M-1)/2} a_{kn} h(n) = H_r(\omega_k) \quad \forall k = 0, 1, \dots, \frac{M-1}{2} \quad (47)$$

donde:

$$a_{kn} = 2 \cos \omega_k \left( \frac{M-1}{2} - n \right) \quad (48)$$

$$a_{kn} = 1 \quad n = \frac{M-1}{2} \quad \forall k \quad (49)$$

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

$$\sum_{n=0}^{(M/2)-1} a_{kn} h(n) = H_r(\omega_k) \quad \forall k = 0, 1, \dots, \frac{M}{2} - 1 \quad (50)$$

# Diseño de un filtro FIR. Determinación de los coeficientes mediante la técnica de muestreo de la frecuencia

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:

$$H_r(0) = 1, \quad H_r(\pi/2) = 1/2 \quad (51)$$

# Diseño de un filtro FIR. Determinación de los coeficientes mediante la técnica de muestreo de la frecuencia

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:

$$\begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix} \begin{bmatrix} h(0) \\ h(1) \end{bmatrix} = \begin{bmatrix} H_r(0) \\ H_r(\pi/2) \end{bmatrix} \quad (52)$$

en este caso:

$$\begin{bmatrix} 2 & 2 \\ -\sqrt{2} & \sqrt{2} \end{bmatrix} \begin{bmatrix} h(0) \\ h(1) \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \end{bmatrix} \quad (53)$$

de donde:  $h(0) = h(3) = 0.0732232$  y  $h(1) = h(2) = 0.4267766$

# Diseño de un filtro FIR. Determinación de los coeficientes mediante la técnica 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 función
%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
    end
    a(k+1,(M-1)/2+1)=1;
end
end h=inv(a)*H;
```

# Diseño de un filtro FIR. Determinación de los coeficientes mediante la técnica de muestreo de la frecuencia

Para diseñar un filtro con  $M = 15$  con la siguiente especificación:

$$H_r\left(\frac{2\pi k}{15}\right) = \begin{cases} 1 & k = 0, 1, 2, 3 \\ 0.4 & k = 4 \\ 0 & k = 5, 6, 7 \end{cases} \quad (54)$$

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
```

# Filtros IIR

- En la ecuación de diferencias 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 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.)



# Filtros IIR

los métodos de diseño están generalmente basados en procedimientos clásicos como:

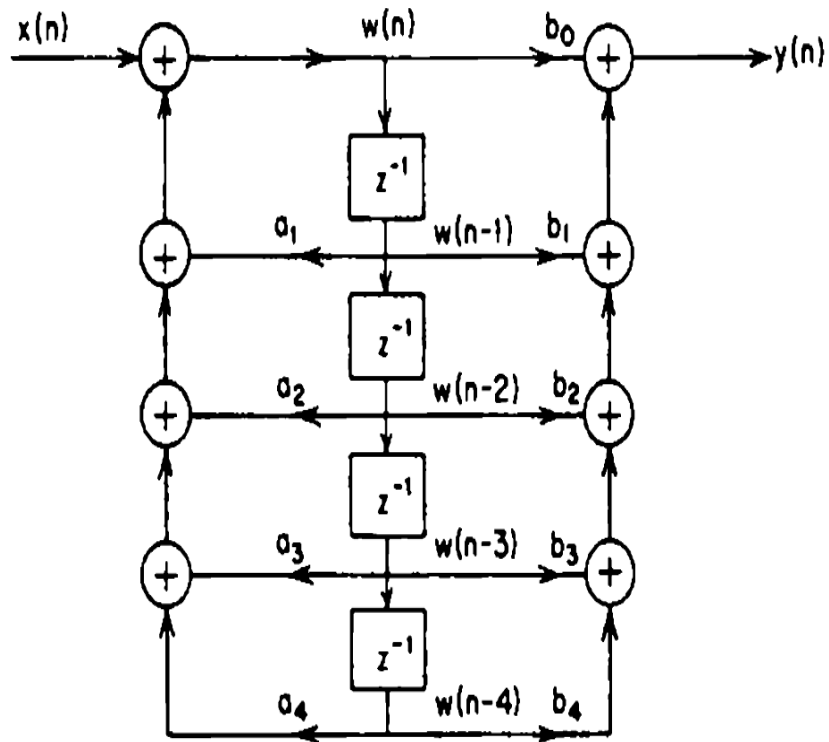
1. Diseño de Butterworth
2. Diseño de Bessel
3. Diseño de Chebyshev
4. Diseños elípticos

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.



# Implementación de filtros IIR con requerimientos mínimos de almacenamiento

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



# 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:

$$e^{aT_s} \quad (55)$$

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:

$$s - a \rightarrow 1 - e^{aT_s} z^{-1} \quad (56)$$

# Diseño de filtros IIR mediante transformación Z apareada

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

$$H_s(s) = \frac{s}{s^2 + 400s + 2 \times 10^5} \quad (57)$$

para una frecuencia de muestreo  $f_s = 1\text{KHz}$ .

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:

$$z = e^{(-200 \pm j400)T_s} = e^{-0.2 \pm j0.4} = 0.7541 \pm j0.3188 \quad (58)$$

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

$$H_z(z) = \frac{1 - z^{-1}}{(1 - (0.7541 + j0.3188)z^{-1})(1 - (0.7541 - j0.3188)z^{-1})} \quad (59)$$

de donde

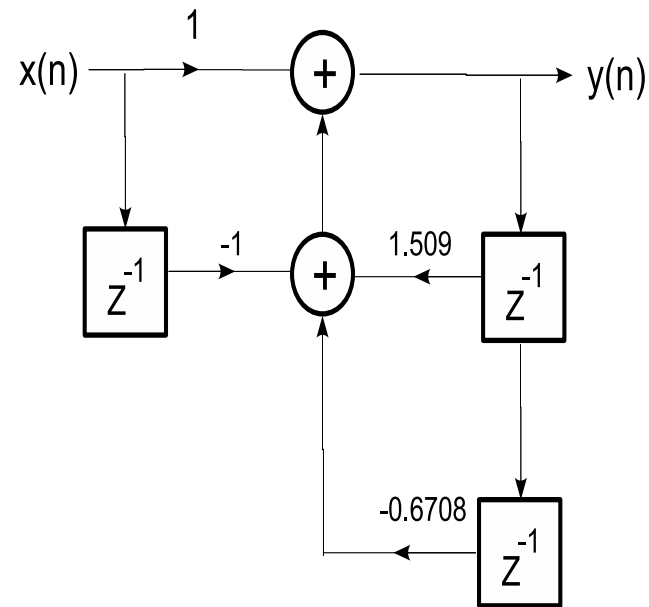
$$H_z(z) = \frac{1 - z^{-1}}{1 - 1.509z^{-1} + 0.6708z^{-2}} \quad (60)$$

# Diseño de filtros IIR mediante transformación Z apareada

La ecuación de diferencias sería entonces:

$$y(n] = x(n) - x(n - 1) + 1.509y(n - 1) - 0.6708y(n - 2) \quad (61)$$

El diagrama de bloques correspondiente sería:



# Diseño de filtros IIR mediante transformación Z apareada

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)
```

BZ y AZ son los coeficientes del polinomio del numerador y del denominador respectivamente de la función sistema de un filtro digital IIR cuya respuesta al impulso es igual a la del filtro analógico cuya función de transferencia en el dominio de Laplace es la 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:

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

BZ =

```
0.0010    -0.0009
```

AZ =

```
1.0000    -1.5082     0.6703
```



# Filtros 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.
- 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.

# Filtros Butterworth

$$|H(\Omega)|^2 = \frac{1}{1 + (\Omega/\Omega_c)^{2N}} \quad (62)$$

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

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

$$\frac{V_o}{V_i} = \frac{G}{\omega_0 s + 1} \quad (63)$$

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

$$\frac{V_o}{V_i} = \frac{G}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2} \quad (64)$$

Los filtros Butterworth solo tienen polos y estos se distribuyen uniformemente a lo largo del perímetro del círculo trigonométrico unitario.

## Muestreo y La regla de Nyquist

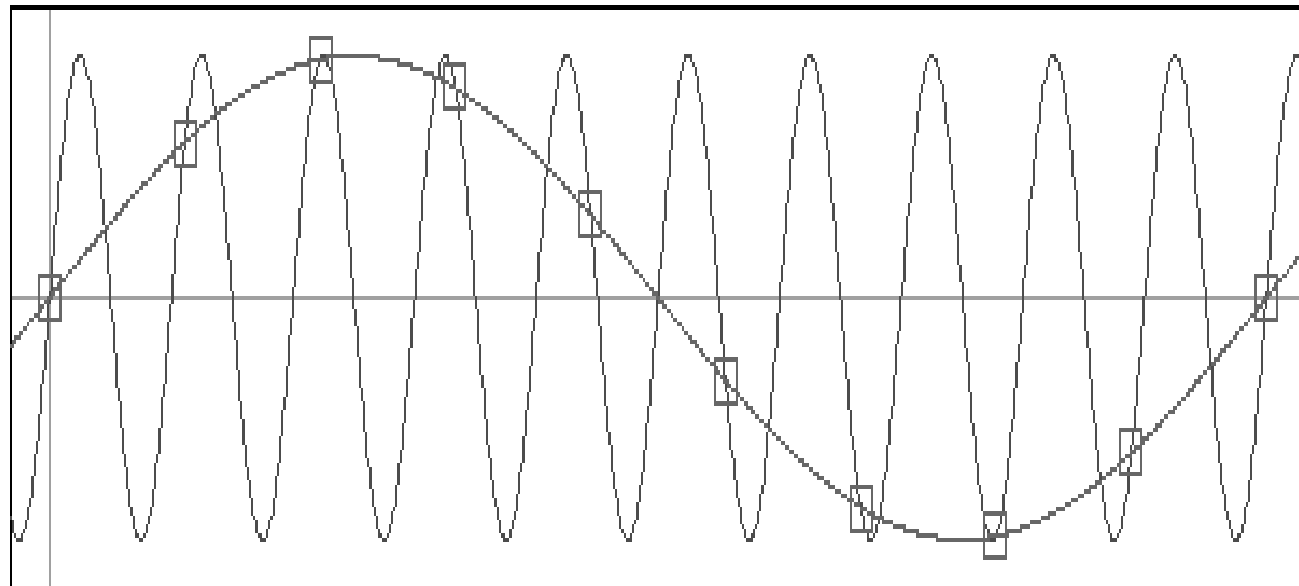
Si una señal  $x_a(t)$  tiene una transformada de Fourier  $X_a(j\omega)$  tal que  $X_a(j\omega) = 0$  para  $\omega \geq 2\pi F_N$ , entonces  $x_a(t)$  puede ser reconstruida a partir de muestras equiespaciadas  $x_a(nT)$  si y solo si  $2F_N < 1/T$ . A la frecuencia  $F_N$  le llamamos la frecuencia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



# El efecto Aliasing

La regla de Nyquist dice que se debe muestrear con una frecuencia de al menos el doble de la frecuencia mayor que aún es de interés, en caso de no cumplir 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





## 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.



## 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 impares, entonces lo denotamos en el diagrama como  $2 \downarrow$ .
- El submuestreo crítico es el 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 más distorsión que un desplazamiento en el tiempo y/o una escalamiento en la amplitud de la señal.

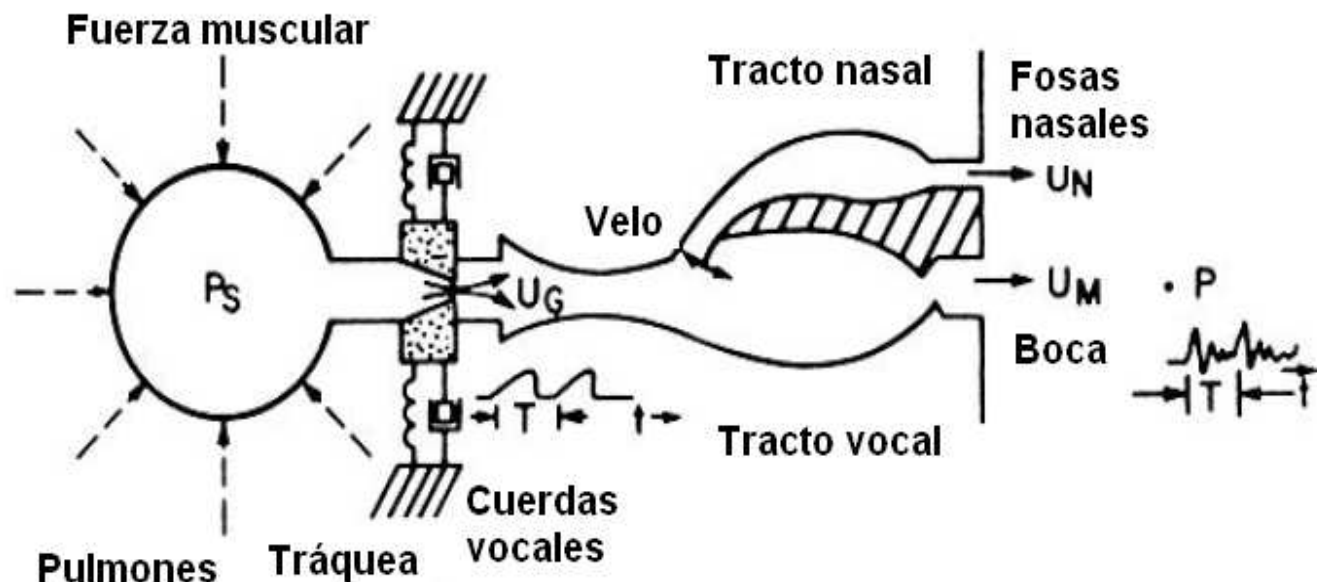


# Modelos digitales de la señal de voz y Teoría acústica

- 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”.
- 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.

# Anatomía y fisiología de la Producción de VOZ

El tracto vocal es como un tubo cuya área de sección transversal no es uniforme. En un extremo están las cuerdas vocales y en el otro los labios. 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.



# Fonética: sonidos vocalizados y no-vocalizados

- 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-periódico debido a la vibración de las cuerdas vocales, a los sonidos producidos de esta manera les llamamos fricativos vocalizados.
- 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.



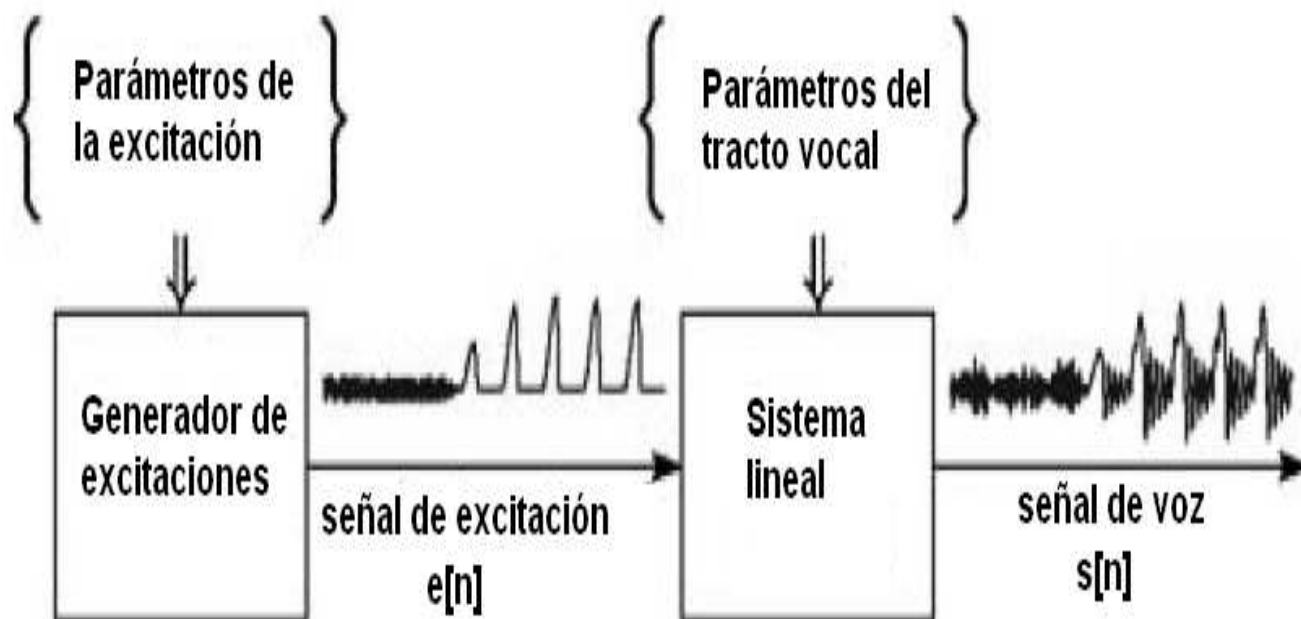
## El concepto de Formantes

- 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”.
- 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 convertirlas en fonemas.





# Técnica de síntesis de sonidos vocalizados usando los primeros tres formantes



# Técnica de síntesis de sonidos vocalizados usando los primeros tres formantes

- Como el tracto vocal cambia relativamente despacio, asumimos que la respuesta del sistema lineal no cambia en menos de 10ms.
- Los coeficientes del filtro  $a_k$  y  $b_k$  cambian a un régimen de 50 a 100 veces por segundo.

$$H(z) = \frac{\sum_{k=0}^M b_k z^{-k}}{1 - \sum_{k=1}^N a_k z^{-k}} \quad (65)$$

- Algunos de los polos 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 de voz, a veces es de utilidad emplear ceros de la función sistema para modelar sonidos fricativos y nasales. Mucho modelos solo emplean polos para simplificar el análisis de estimación de los parámetros.
- 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 son constantes por intervalos de tiempo cortos.

# Teoría Acústica

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

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial(u/A)}{\partial t} \quad (66)$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2} \frac{\partial(pA)}{\partial t} + \frac{\partial A}{\partial t} \quad (67)$$

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  $t$ .

$\rho$  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  $t$ .



# Teoría Acústica

- 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”
- 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.
- La función de área  $A(x, t)$  debe ser conocida y como es de imaginarse, mediciones detalladas son muy difíciles de obtener
- Hay quien lo ha intentado a partir de secuencias de imágenes de rayos X.

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

El caso mas simple, 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, las ecuaciones se reducen a:

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A} \frac{\partial u}{\partial t} \quad (68)$$

$$-\frac{\partial u}{\partial x} = \frac{A}{\rho c^2} \frac{\partial p}{\partial t} + \frac{\partial A}{\partial t} \quad (69)$$

La solución de este sistema de ecuaciones diferenciales tiene la forma:

$$u(x, t) = [u^+(t - x/c) - u^-(t + x/c)] \quad (70)$$

$$p(x, t) = \frac{\rho c}{A} [u^+(t - x/c) + u^-(t + x/c)] \quad (71)$$

$u^+(t - x/c)$  y  $u^-(t + x/c)$  son ondas viajando en la dirección pos y neg respectivamente.

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

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:

$$-\frac{\partial v}{\partial x} = L \frac{\partial i}{\partial t} \quad (72)$$

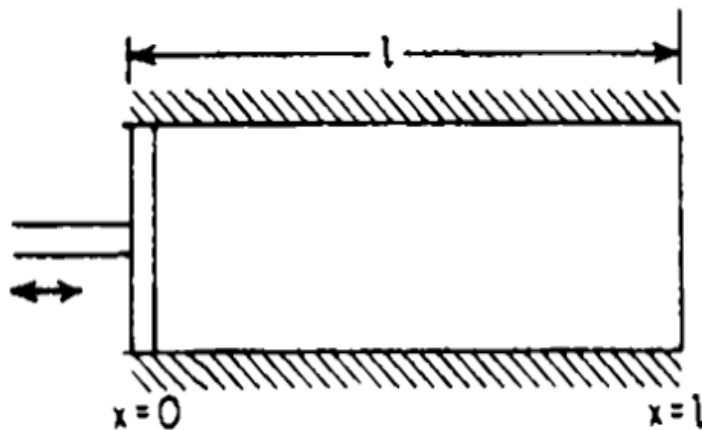
$$-\frac{\partial i}{\partial x} = C \frac{\partial v}{\partial t} \quad (73)$$

$L$  y  $C$  son la inductancia y la capacitancia por unidad de longitud respectivamente

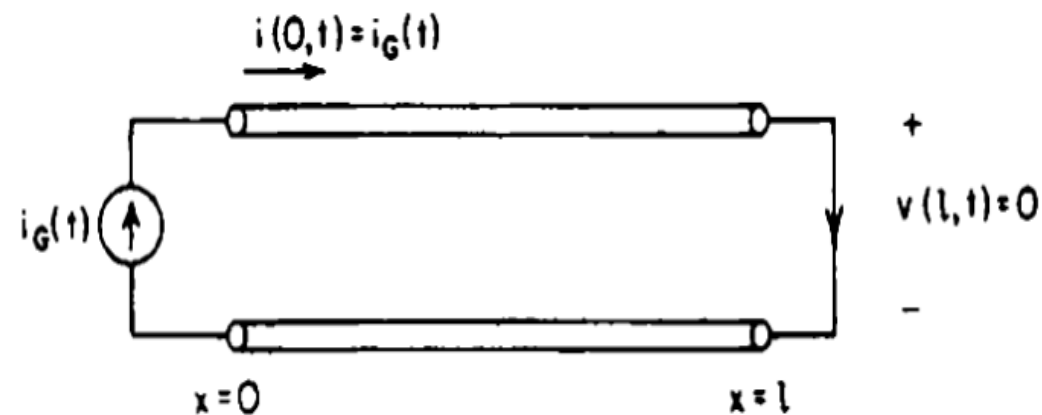
Variable acústica	Variable Eléctrica
$p$ presión	$v$ voltaje
$u$ velocidad de volumen	$i$ corriente
$\rho/A$ inductancia acústica	$L$ inductancia
$A/\rho c^2$ Capacitancia acústica	$C$ capacitancia

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

El tubo de sección transversal uniforme al que un pistón inyecta aire se comporta en forma idéntica a una línea de transmisión terminada en cortocircuito ( $V(l, t) = 0$ ) y excitada por una fuente de corriente.



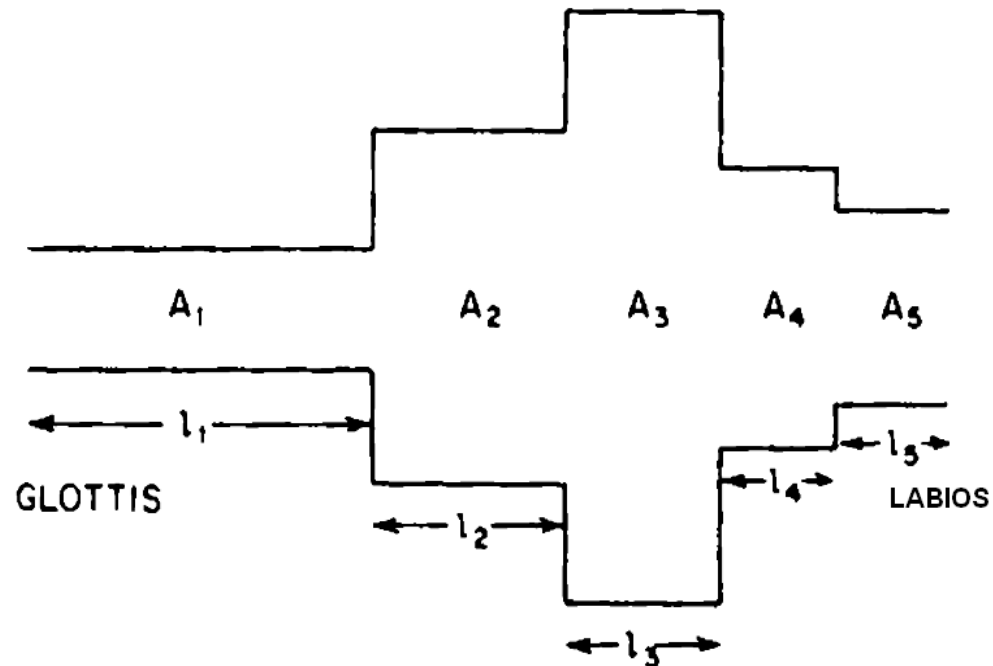
(a)



(b)

# Modelo del tracto vocal basado en tubos sin pérdidas de sección uniforme 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





# Modelo del tracto vocal basado en tubos sin pérdidas de sección uniforme concatenados

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:

$$u_k(x, t) = [u_k^+(t - x/c) - u_k^-(t + x/c)] \quad (74)$$

$$p_k(x, t) = \frac{\rho c}{A_k} [u_k^+(t - x/c) + u_k^-(t + x/c)] \quad (75)$$

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

# 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:

$$p_k(l_k, t) = p_{k+1}(0, t) \quad (76)$$

$$u_k(l_k, t) = u_{k+1}(0, t) \quad (77)$$

De donde:

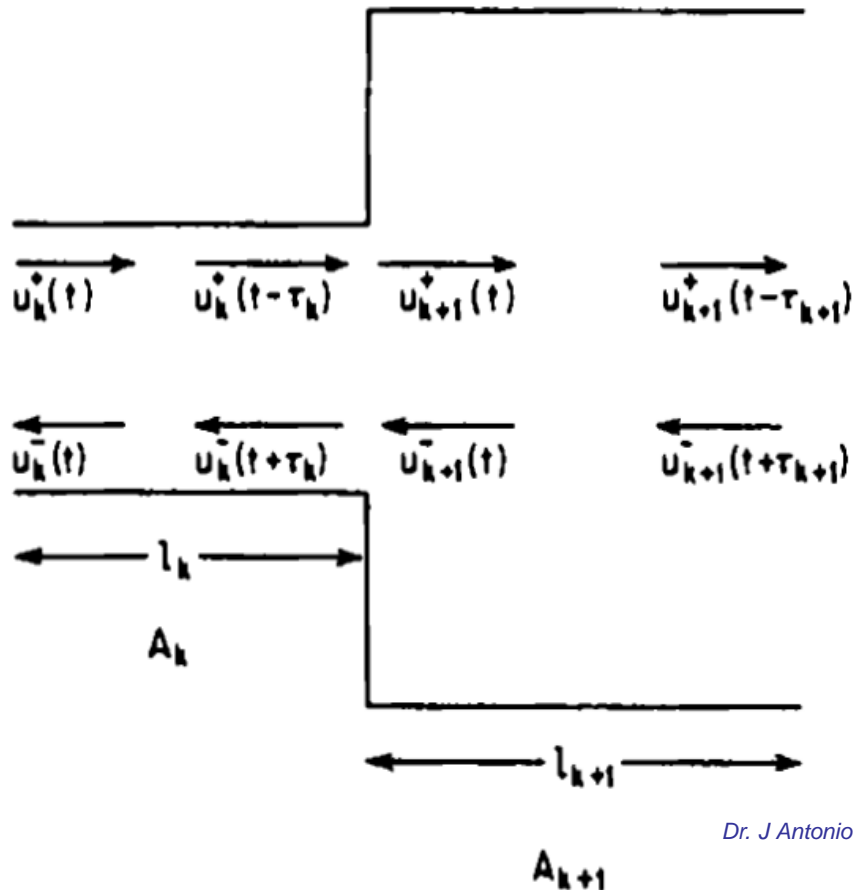
$$\frac{A_{k+1}}{A_k} [u_k^+(t - \tau_k) + u_k^-(t + \tau_k)] = [u_{k+1}^+(t) + u_{k+1}^-(t)] \quad (78)$$

$$u_k^+(t - \tau_k) - u_k^-(t + \tau_k) = u_{k+1}^+(t) + u_{k+1}^-(t) \quad (79)$$

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

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

$u_{k+1}^+(t)$  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  $k$ .



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

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  $u_k^-(t + \tau_k)$  y sustituyendo obtenemos:

$$u_{k+1}^+(t) = \frac{2A_{k+1}}{A_{k+1} + A_k} u_k^+(t - \tau_k) + \frac{A_{k+1} - A_k}{A_{k+1} + A_k} u_{k+1}^-(t) \quad (80)$$

Ahora, restando miembro a miembro obtenemos:

$$u_k^-(t + \tau_k) = -\frac{A_{k+1} - A_k}{A_{k+1} + A_k} u_k^+(t - \tau_k) + \frac{2A_k}{A_{k+1} + A_k} u_{k+1}^-(t) \quad (81)$$

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

La cantidad:

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \quad (82)$$

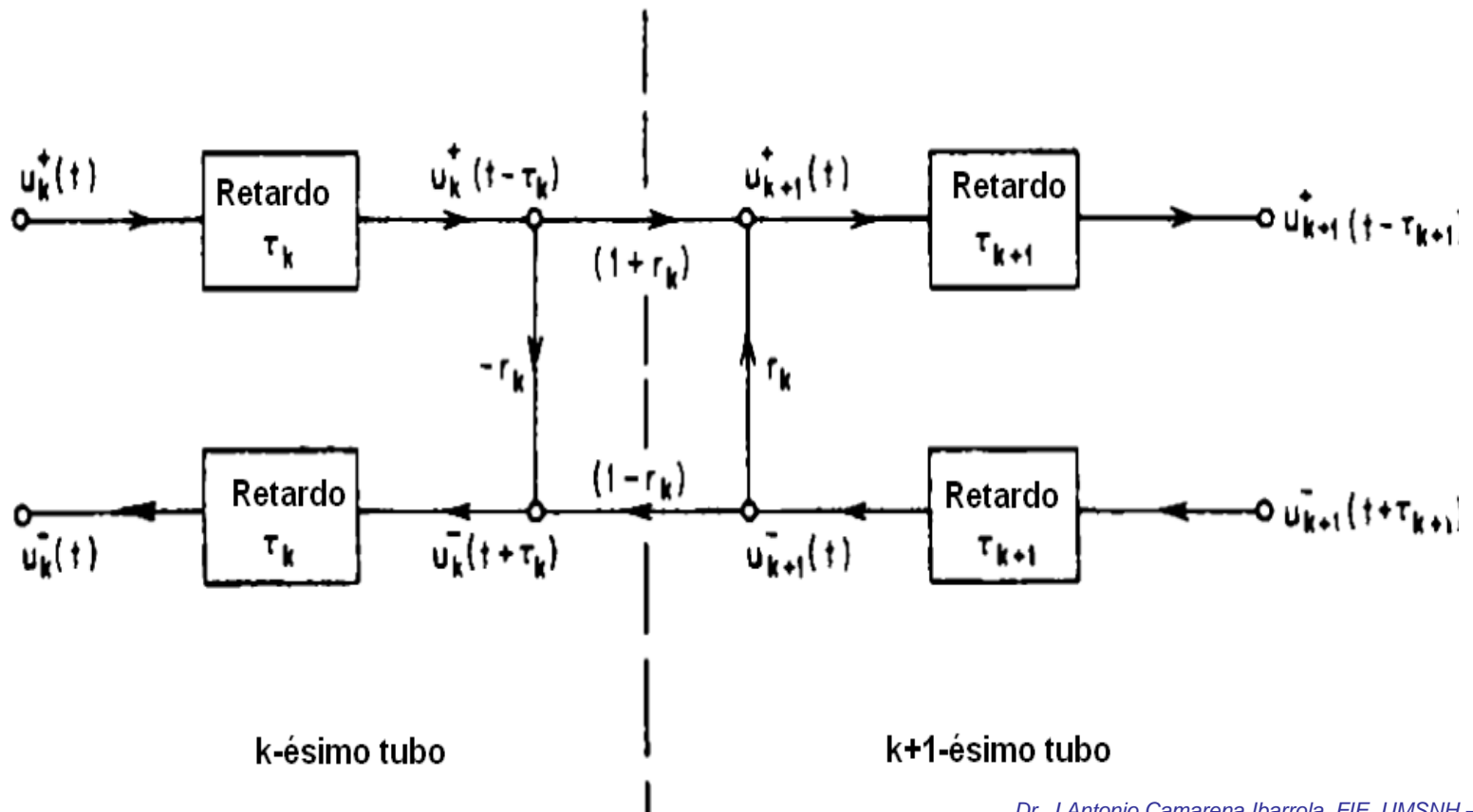
es la fracción de  $u_k^-(t)$  que es reflejada en la unión, por lo cual se le denomina “coeficiente de reflexión de la k-ésima unión”. En vista de que todas las áreas de sección transversal son necesariamente positivas se puede demostrar que  $-1 < r_k < 1$ . Las ecuaciones de propagación de las ondas 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.

$$u_{k+1}^+(t) = (1 + r_k)u_k^+(t - \tau_k) + r_k u_{k+1}^-(t) \quad (83)$$

$$u_k^-(t + \tau_k) = -r_k u_k^+(t - \tau_k) + (1 - r_k)u_{k+1}^-(t) \quad (84)$$

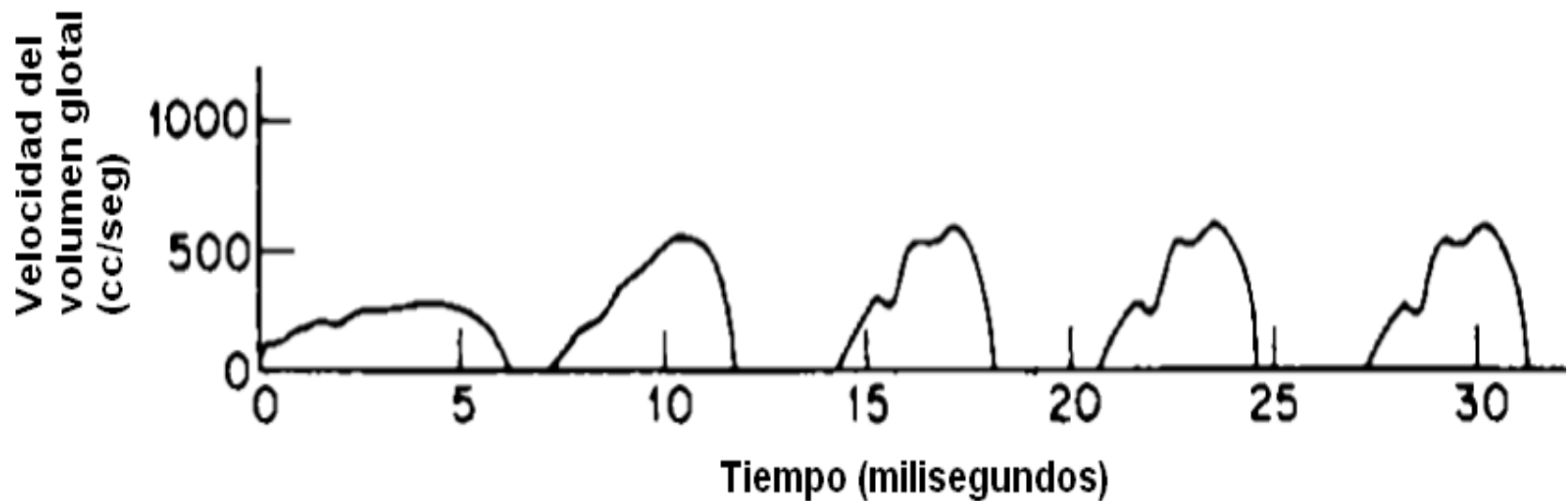
# Síntesis de voz usando coeficientes reflejantes

Kelly and Lochbaum usaron estas ecuaciones para hacer síntesis de voz.



## 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:



## Pulso glotal

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:

$$g(n) = \frac{1}{2} [1 - \cos(\pi n / N_1)] \quad 0 \leq n \leq N_1 \quad (85)$$

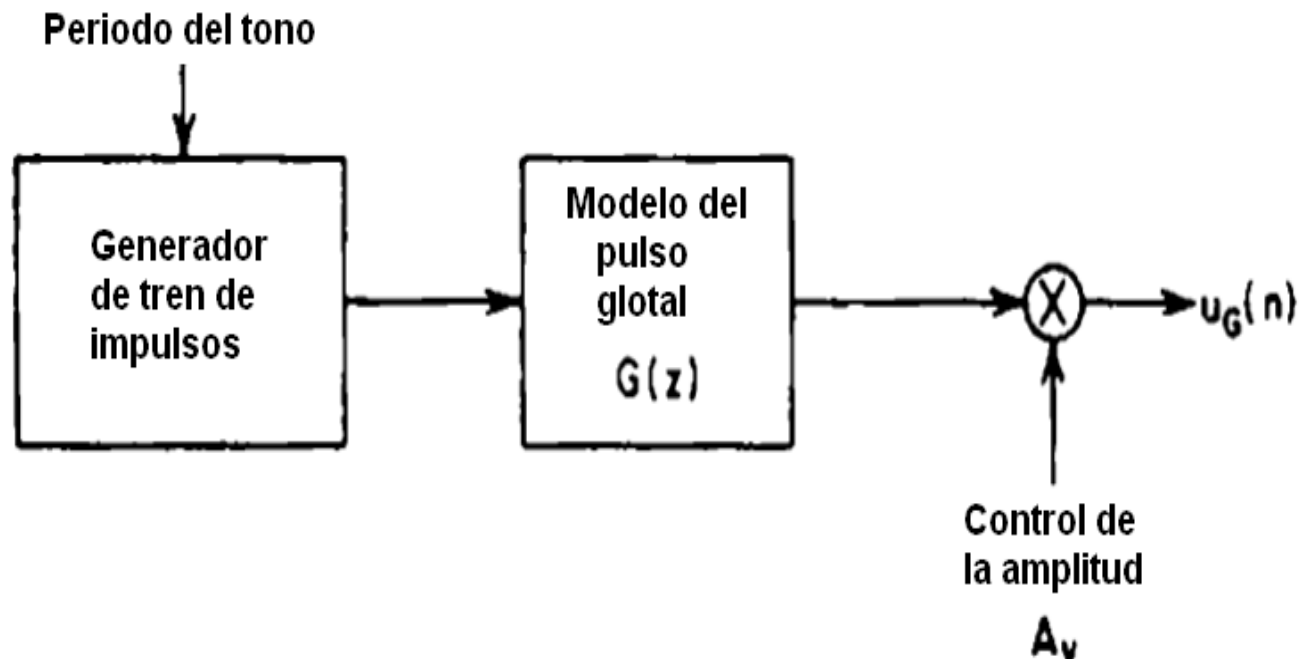
$$= \cos(\pi(n - N_1) / 2N_2) \quad N_1 \leq n \leq N_1 + N_2 \quad (86)$$

$$= 0 \quad \text{en otro lugar} \quad (87)$$



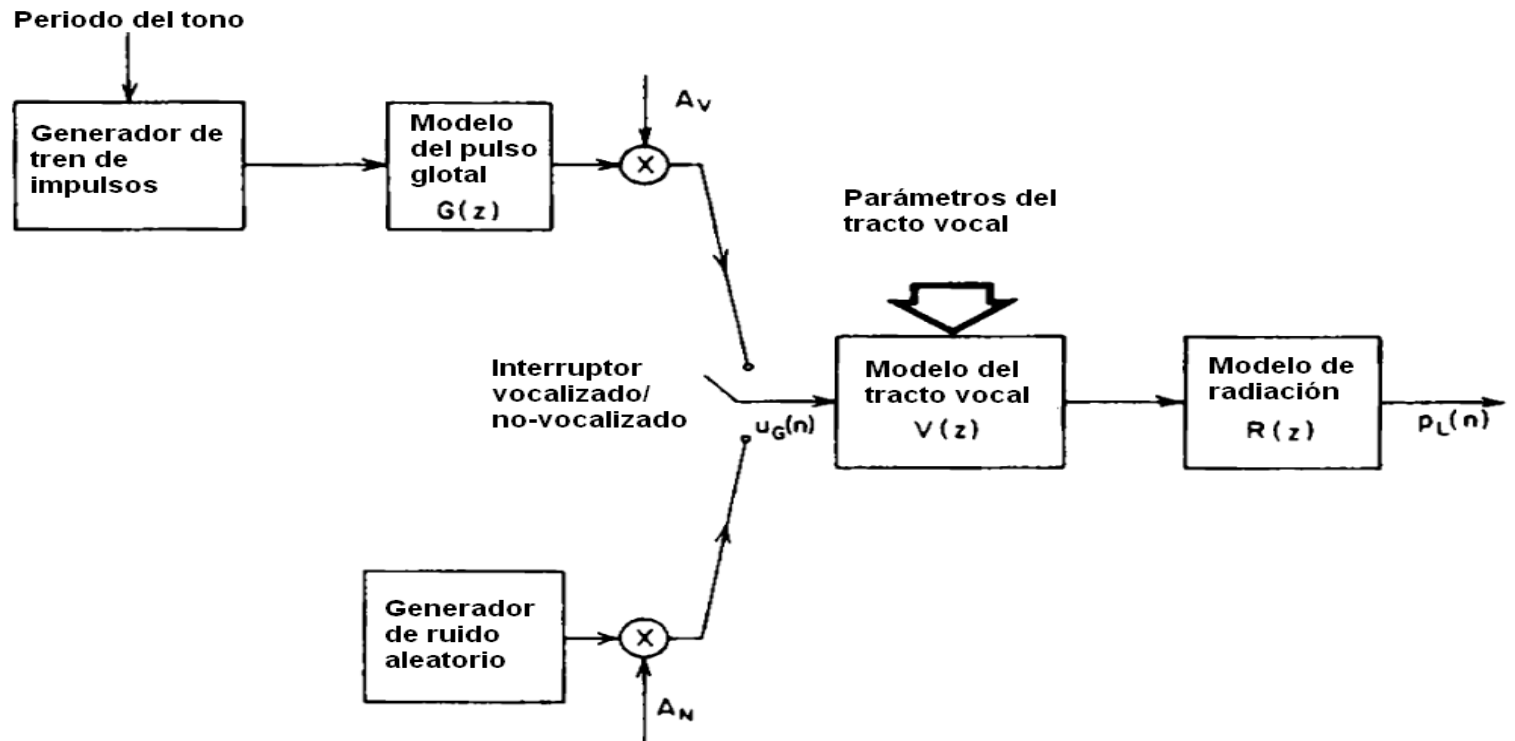
## Pulso glotal

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  $G(z)$ .



# Modelo completo del tracto vocal

Poniendo todos los ingredientes juntos





## Modelo completo del tracto vocal

- 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.
- 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.



# Procesamiento de la señal de voz en el dominio del tiempo

- 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.
- Energía de tiempo corto.
- Régimen de cruces por cero de tiempo corto
- Autocorrelación de tiempo corto
- Entropía de tiempo corto
- 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)
- En caso de tratarse de sonidos vocalizados determinar el tono (autocorrelación modificada)
- Se puede realizar reconocimiento de voz con características extraídas en el dominio del tiempo (usando cruces por cero de orden superior).

## Energía de tiempo corto

La energía  $E$  de una señal  $x$  se define como:

$$E = \sum_{m=-\infty}^{\infty} x^2(m) \quad (88)$$

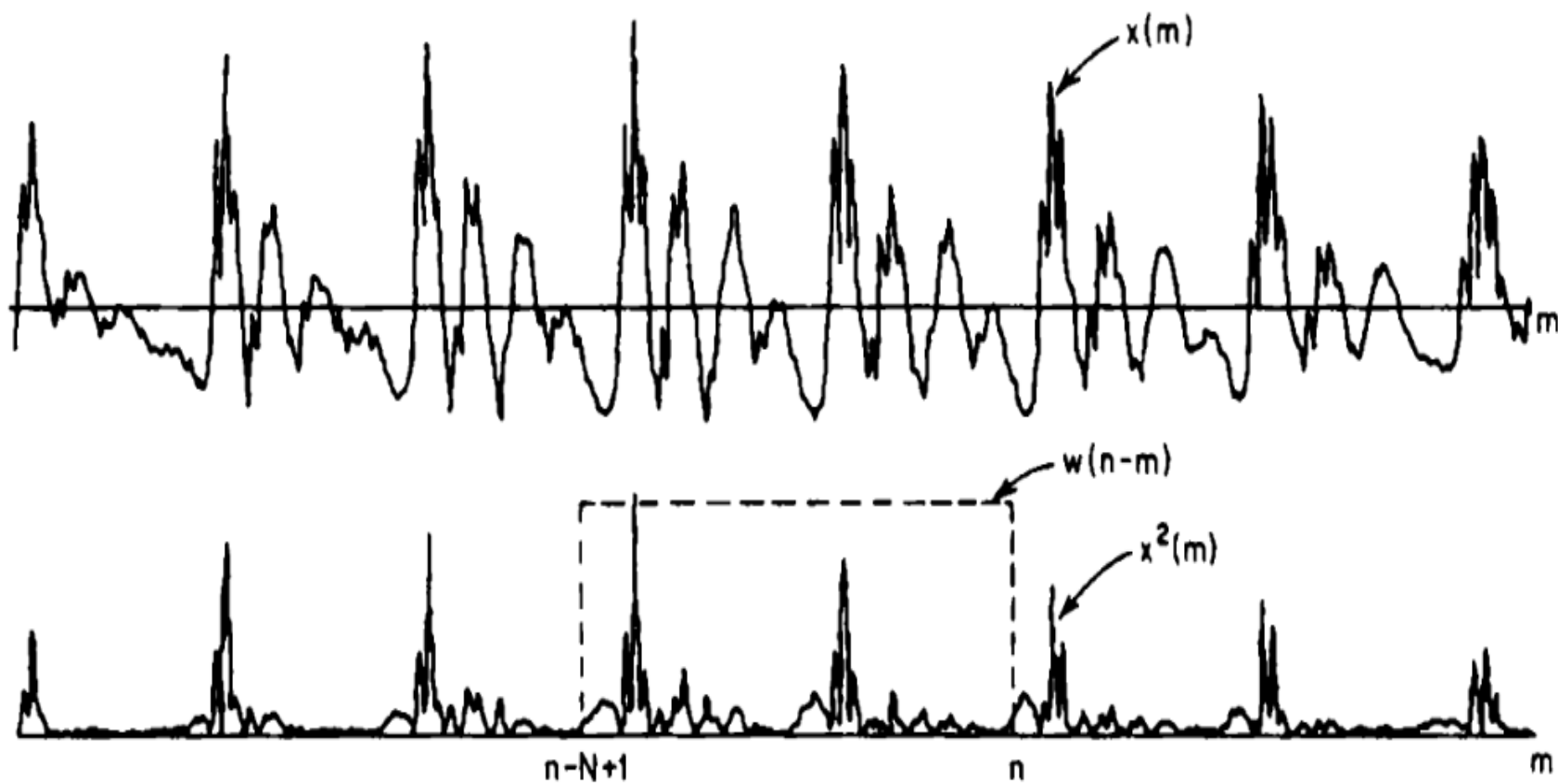
Lo mas parecido a la energía instantánea es la denominada *energía de tiempo corto*  $E_n$  que es la energía contenida en un segmento de la señal que comienza en el tiempo discreto  $n$  y se determina mediante:

$$E_n = \sum_{m=n-N+1}^n x^2(m) \quad (89)$$

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 rectangular es:

$$w(n) = \begin{cases} 1 & 0 \leq n \leq N - 1 \\ 0 & \text{en otro lado} \end{cases} \quad (90)$$

# Energía de tiempo corto



## Energía de tiempo corto

La señal  $x(m)$  se eleva al cuadrado y luego se ve a través de una ventana de tamaño  $N$ , de manera que las 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:

$$E_n = \sum_{m=-\infty}^{m=\infty} [x(m)w(n-m)]^2 \quad (91)$$

o bien mediante:

$$E_n = \sum_{m=-\infty}^{m=\infty} [x(m)^2 h(n-m)] \quad (92)$$

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



## Magnitud de tiempo corto

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

$$M_n = \sum_{m=-\infty}^{m=\infty} [x(m)w(n - m)] \quad (93)$$



## 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. Una definición adecuada del régimen de cruces por cero de tiempo corto es:

$$Z_n = \sum_{m=-\infty}^{m=\infty} [|\text{signo}[x(m)] - \text{signo}[x(m-1)]|w(n-m)] \quad (94)$$

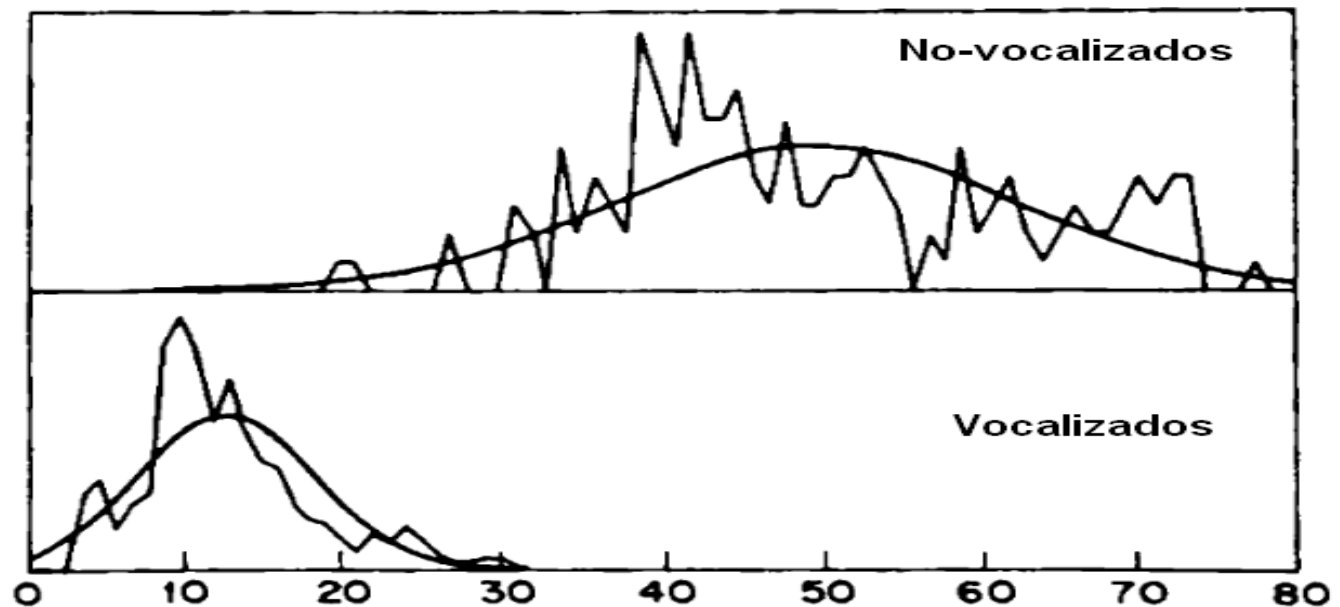
donde:

$$\text{signo}[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases} \quad (95)$$

además:

$$w(n) = \begin{cases} \frac{1}{2N} & 0 \leq n \leq N-1 \\ 0 & \text{en otro lado} \end{cases} \quad (96)$$

## Régimen de cruces por cero de tiempo corto



Un uso que podemos dar al régimen de cruces por cero de tiempo corto es el de discriminar sonidos vocalizados de sonidos no-vocalizados. 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.



## 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 convirtiéndola 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



## Estimación del tono

1. Pasar la señal por un filtro pasabajas con una frecuencia de corte de 900Hz
2. Localizar picos y valles 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)



## Estimación del tono

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  $\tau$  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)

## 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  $v_1, v_2, \dots, v_n$  los posibles 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  $p_1, p_2, \dots, p_n$  es la función de Distribución de Probabilidades (PDF por sus siglas en inglés) sujeta a:

$$\sum_{i=1}^n p_i = 1 \quad (97)$$

## Entropía de la señal de voz

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  $I(p_i)$ .

$I(p_i)$  debe de ser calculada de manera que si  $v_i$  depende de dos o más eventos independientes con probabilidades  $p_{i1}, p_{i2}, \dots$ , 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  $p = p_{i1}p_{i2}, \dots$ , entonces,  $I(p_i) = I(p_{i1}) + I(p_{i2}), \dots$ . 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

$$I(p_i) = \ln\left(\frac{1}{p_i}\right) = -\ln(p_i) \quad (98)$$

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.

## Entropía de la señal de voz

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. Al equivalente continuo se le conoce como “Entropía diferencial”. Como la 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. 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  $\log(n)$

$$H(x) = E[I(p)] = \sum_{i=1}^n p_i I(p) = - \sum_{i=1}^n p_i \ln(p_i) \quad (99)$$



## Entropía de la señal de voz

$$H_{min} = - \sum_i \delta(k) \ln[\delta(k)] = -\ln(1) = 0 \quad (100)$$

$$H_{max} = - \sum_i \frac{1}{n} \ln\left(\frac{1}{n}\right) = -\ln\left(\frac{1}{n}\right) = \ln(n) \quad (101)$$

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.



# Entropía de la señal de voz

- Para calcular la entropía de una señal se debe hacer una estimación de la función de distribución de probabilidades.
- Si se opta por usar un histograma se deben involucrar suficientes muestras para lograr una estimación confiable
- 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.
- En los métodos paramétricos, primero se debe asumir el tipo de distribución al que se apegará la señal, luego se deben estimar sus parámetros.
- En los métodos no paramétricos no es necesario asumir nada respecto al tipo de distribución; la función de distribución es moldeada por los datos y suavizada por un kernel en un proceso iterativo
- El más popular de estos métodos es el de la ventana de Parzen, sin embargo, los métodos no paramétricos son costosos computacionalmente y por ende no muy usados en aplicaciones de tiempo real.

## Entropía de la señal de voz

La entropía de una variable aleatoria con distribución  $N(0, R)$ .

$$p_i = \frac{f_i}{N} \quad (102)$$

donde  $f_i$  es el número de veces que el valor  $v_i$  ocurre en la señal  $x$

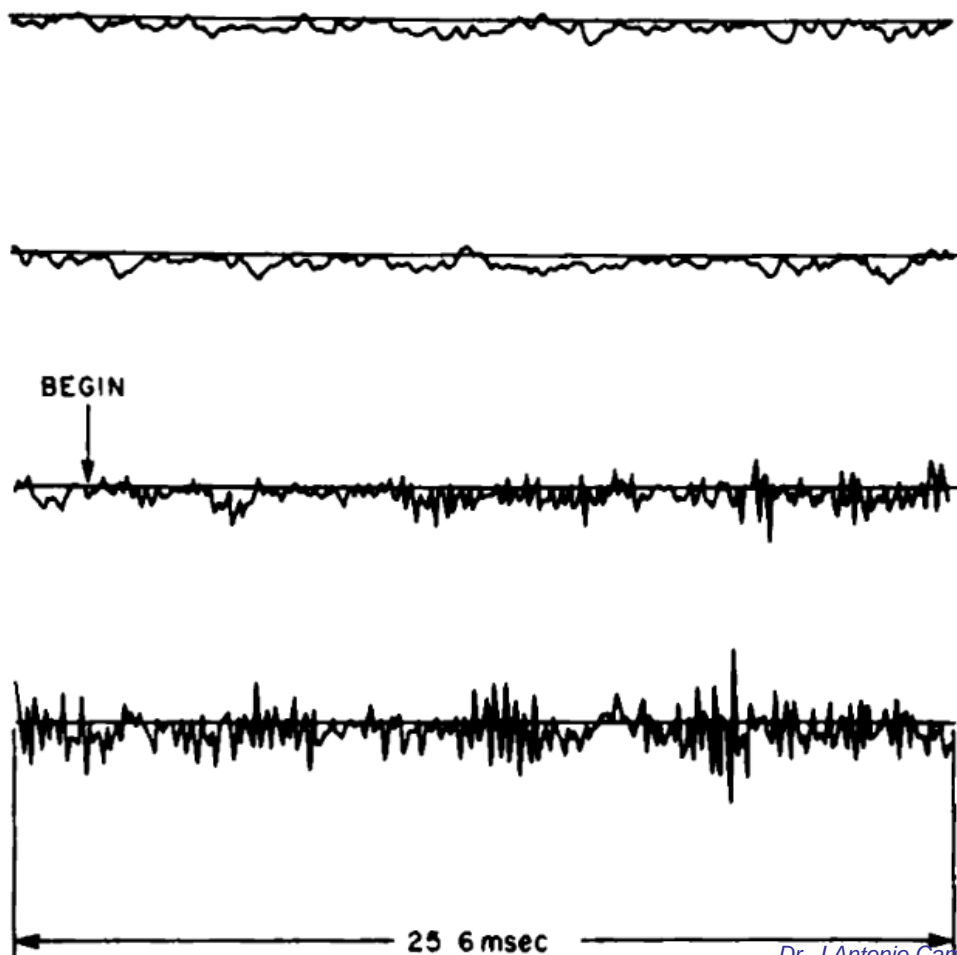
$$f_i = \sum_{j=1}^N \varphi(x_j, v_i) \quad (103)$$

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

$$H = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln(|R|) \quad (104)$$

donde  $R$  es la matriz de covarianzas de grado  $n$ .

# Segmentación de palabras aisladas





## Segmentación de palabras aisladas

- 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
- 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
- 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).

# Autocorrelación de tiempo corto

La Autocorrelación de una señal  $x$  es una función definida como:

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m-k) \quad (105)$$

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:

$$\phi(k) = \phi(k + P) \quad (106)$$

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  $\phi(0)$

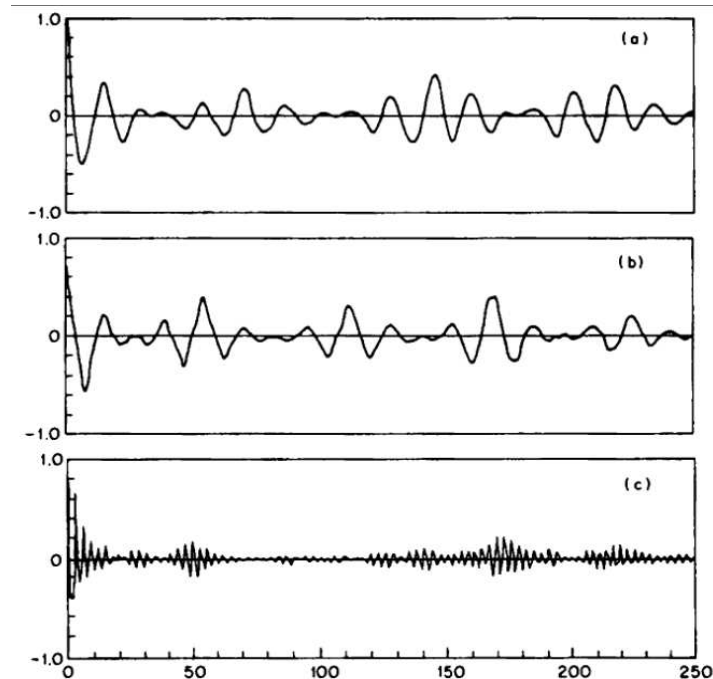
# Autocorrelación de tiempo corto

- La función de autocorrelación tiene máximos locales en  $0, \pm P, \pm 2P, \dots$ ,
- El periodo de la señal puede ser determinado encontrando la ubicación del primer máximo local de la función de autocorrelación.
- La función de autocorrelación es útil 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 determinar los coeficientes LPC utilizados para comprimir y descomprimir la señal.

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)x(m+k)w(n-k-m) \quad (107)$$

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:

## Autocorrelación de tiempo corto



Frecuencia de muestreo de 10KHz;  $N = 401$ . (a) y (b) vocalizados y (c) no-vocalizado. En (a) los picos ocurren aproximadamente en múltiplos de 72 indicando un periodo de 7.2 ms, (frecuencia fundamental de 140 Hz), en (b) los picos ocurren en múltiplos de 58 muestras indicando un periodo de 5.8 ms. En (c) no hay picos periódicos indicando ausencia de periodicidad.

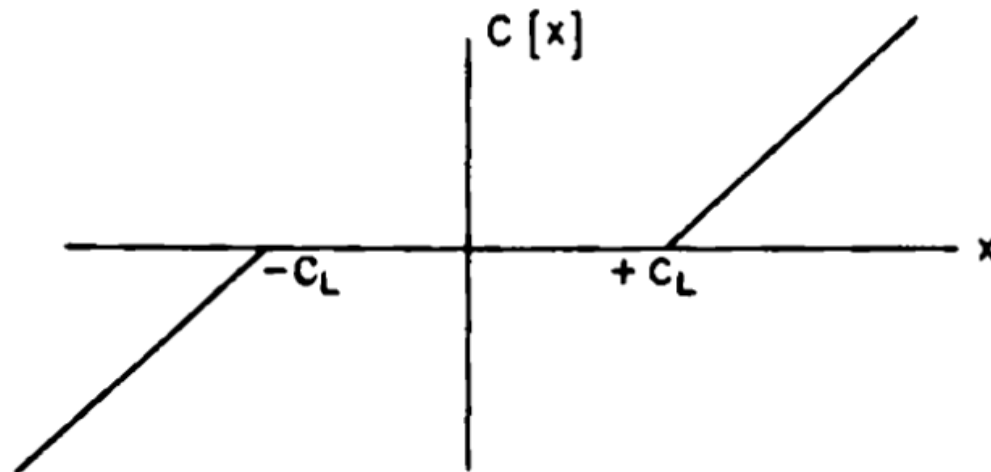


# Algoritmo de discriminación silencio/voz vocalizado/no-vocalizado y estimación del tono mediante autocorrelación

- 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.
- 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*

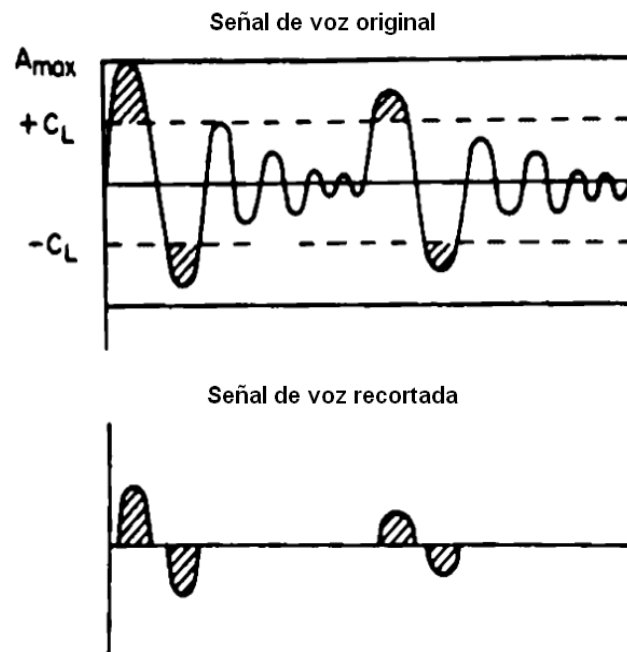
# Algoritmo de discriminación silencio/voz vocalizado/no-vocalizado y estimación del tono mediante autocorrelación

$$y(n) = C[x(n)]$$



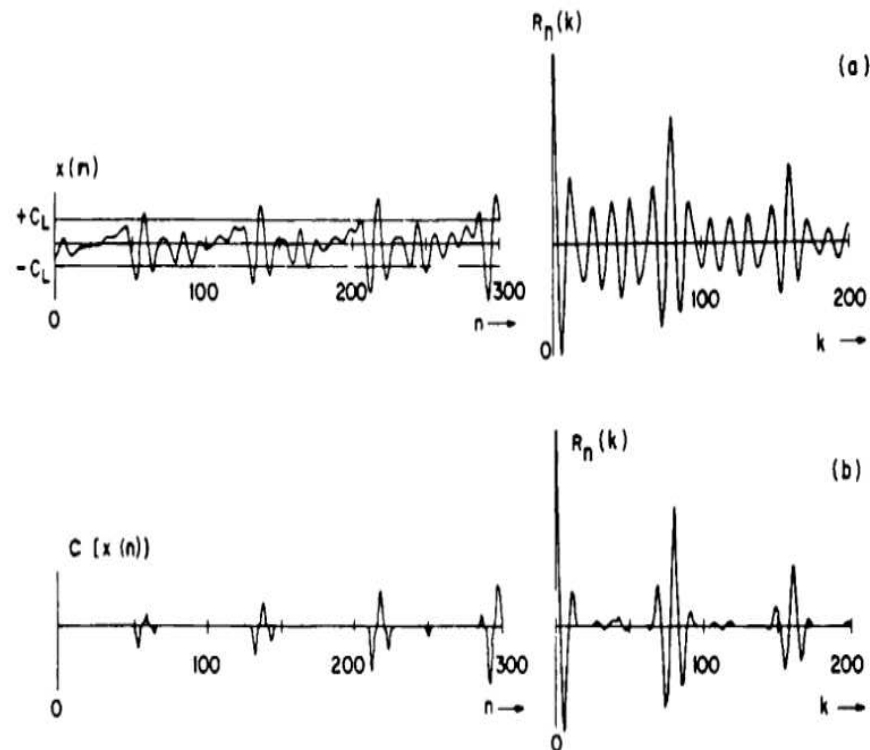
# Algoritmo de discriminación silencio/voz vocalizado/no-vocalizado y estimación del tono mediante autocorrelación

En la Figura 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.

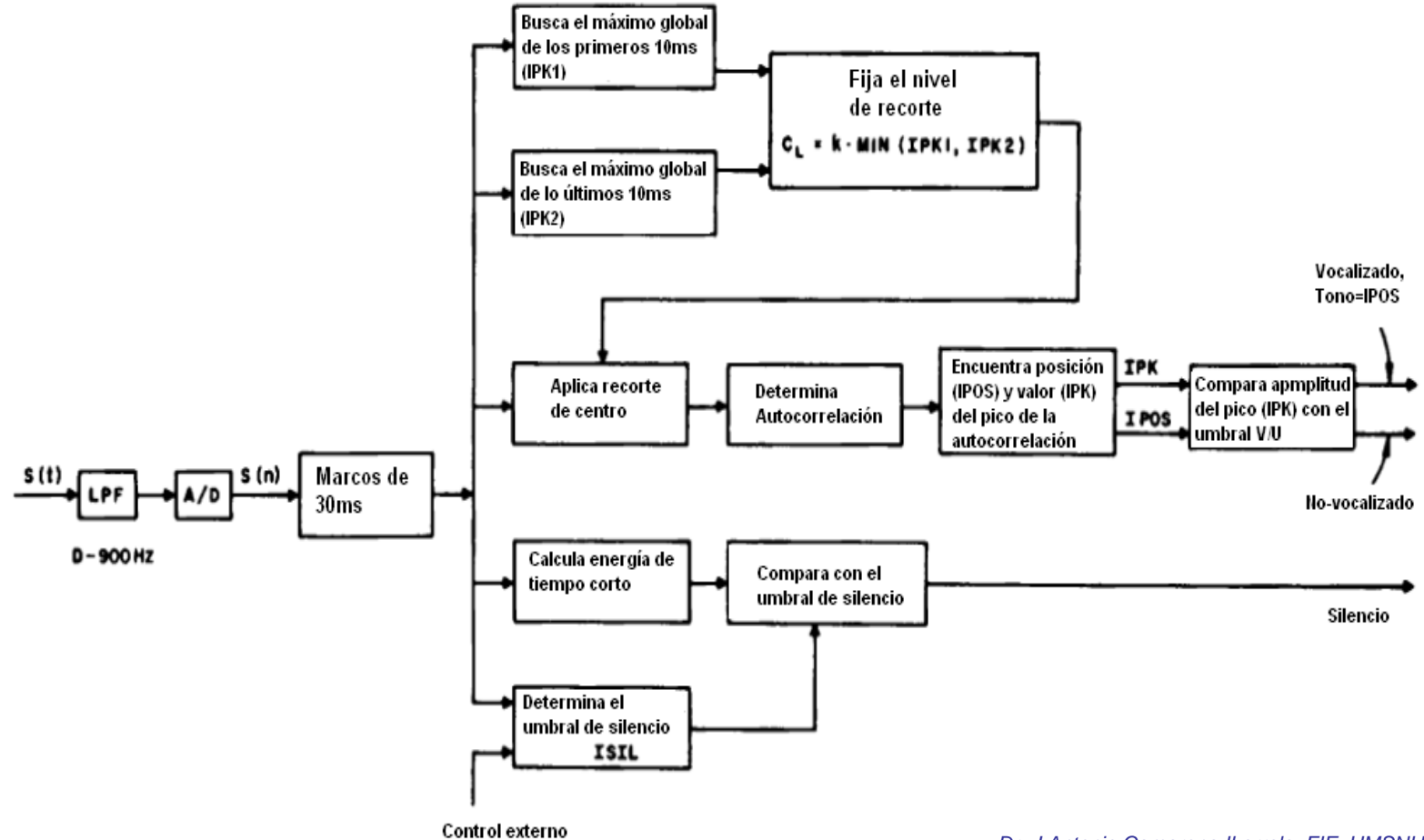


# Algoritmo de discriminación silencio/voz vocalizado/no-vocalizado y estimación del tono mediante autocorrelación

En (a) vemos un sonido vocalizado y su función de autocorrelación y en (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.

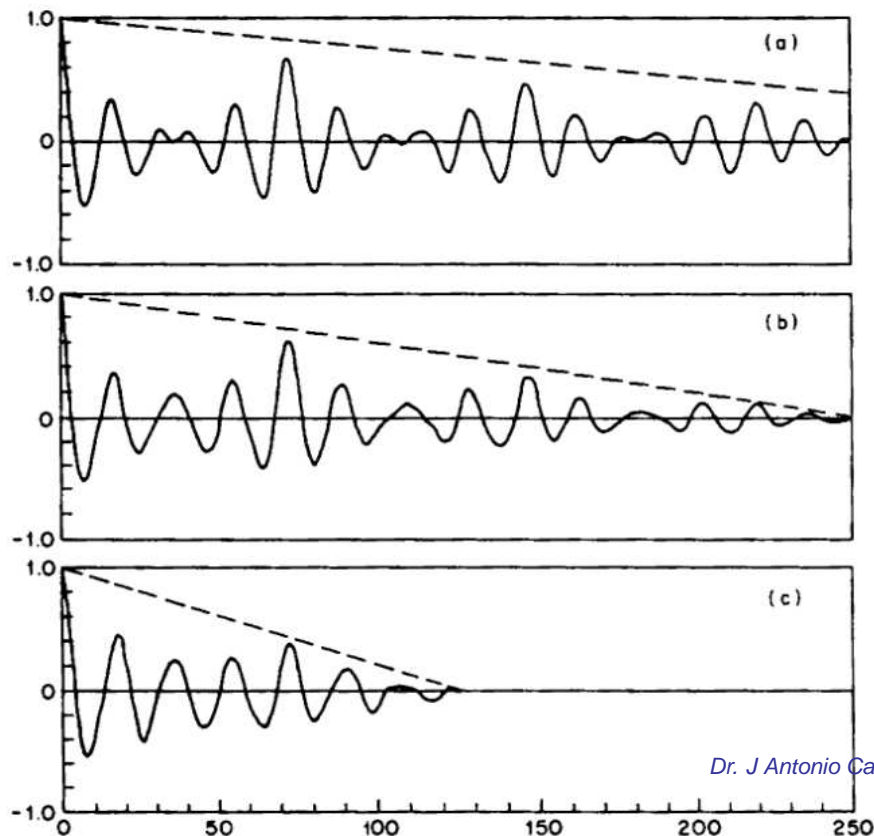


# Algoritmo de discriminación silencio/voz vocalizado/no-vocalizado y estimación del tono mediante autocorrelación



# 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 una reducción en la amplitud de los picos a medida que  $k$  aumenta



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

La función de autocorrelación modificada se define como:

$$\hat{R}_n(k) = \sum_{m=-\infty}^{\infty} x(m)w_1(n-m)x(m+k)w_2(n-k-m) \quad (109)$$

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

$$w_1(m) = \begin{cases} 1 & 0 \leq m \leq N - 1 \\ 0 & \text{otro lugar} \end{cases} \quad (110)$$

$$w_2(m) = \begin{cases} 1 & 0 \leq m \leq N - 1 + K \\ 0 & \text{otro lugar} \end{cases} \quad (111)$$

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

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:

$$\hat{R}_n(k) = \sum_{m=-\infty}^{\infty} x(n+m)w_1(m)x(n+m+k)w_2(m+k) \quad (112)$$

Para la ventana rectangular, esta definición es equivalente a:

$$\hat{R}_n(k) = \sum_{m=0}^{N-1} x(n+m)x(n+m+k) \quad \forall 0 \leq k \leq K \quad (113)$$

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  $x(n+m)w_2(m)$



# 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  $k = 1$ :

$$\hat{R}_n(1) = \sum_{m=0}^{N-1} x(m+n)x(m+n+1) \quad (114)$$

$$= x(n)x(n+1) + x(n+1)x(n+2) + \dots + x(n+N-1)x(n+N) \quad (115)$$

$$= x(n+1)[x(n) + x(n+2)] + x(n+3)[x(n+2) + x(n+4)] + \dots \quad (116)$$

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

$$\hat{R}(k) = B(k) + C(k) \quad (117)$$

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

$$N = 2qk + ak + b \quad (118)$$

donde  $a = 0$ , o bien  $a = 1$  y  $b$  está en el rango  $0 \leq b < k$ . Dr. J Antonio Camarena Ibarrola, FIE, UMSNH – p.105/21

# Algoritmo de Blankenship para el cálculo eficiente de la autocorrelación

$$B(k) = \sum_{j=0}^{q-1} \sum_{i=1}^k x(2jk + i + k)[x(2jk + i) + x(2jk + i + 2k)] \quad (119)$$

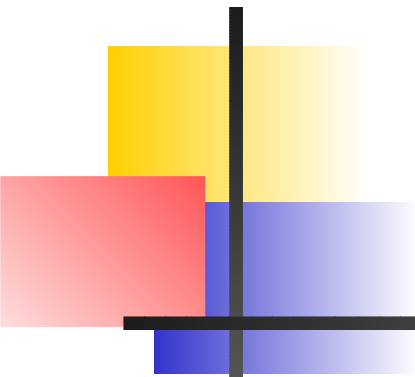
si  $a = 0$ :

$$C(k) = \sum_{i=1}^b x(2qk + i)x(2qk + i + k) \quad (120)$$

y si  $a = 1$

$$C(k) = \sum_{i=1}^b x(2qk + i + k)[x(2qk + i) + x(2qk + i + 2k)] + \sum_{i=b+1}^k x(2qk + i)x(2qk + i + k) \quad (121)$$

El método de Blankenship es particularmente útil cuando solo se requieren algunos valores de la autocorrelación. Si se requieren todos los valores de la autocorrelación conviene hacerlo mediante la transformada discreta inversa de Fourier del espectro de potencia (el cuadrado del espectro de magnitudes)



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

- El reconocimiento de voz, ya sea de palabras aisladas o de habla continua difícilmente se puede hacer en el dominio del tiempo
- 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.



## Transformada de Fourier de tiempo corto

- Necesitamos un análisis espectral 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 significativamente.
- 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 sucesivamente, 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
- Tendremos información de la señal en el dominio de la frecuencia pero sin perder el dominio del tiempo.
- Es común Avanzar los marcos en 10ms cada vez



## 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
- Aplicar la ventana rectangular es equivalente a no aplicar ninguna ventana.

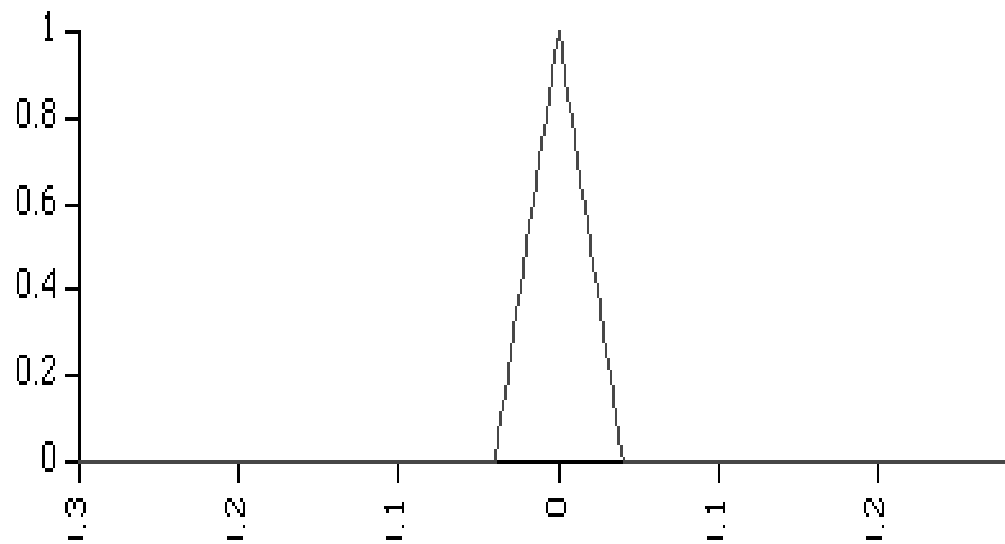
## La ventana de Welch

$$welch(n) = 1 - \left( \frac{n - N/2}{N/2} \right)^2 \quad (122)$$



## La ventana de Barlett

$$\text{barlett}(n) = \begin{cases} 1 - 2n/N & 0 \leq n \leq N/2 \\ 1 + 2n/N & -N/2 \leq n \leq 0 \\ 0 & \text{otro lugar} \end{cases} \quad (123)$$

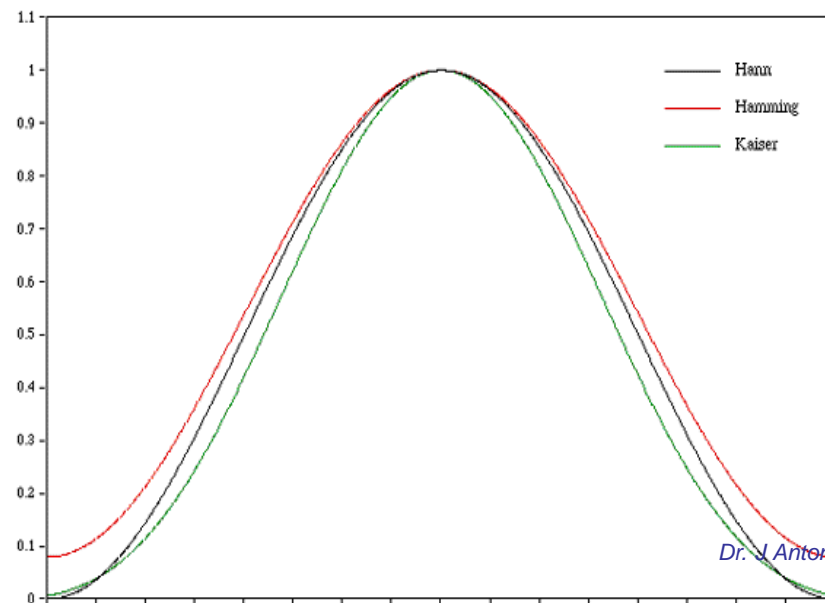


# La ventana de Hamming y la ventana de Hann

$$\text{hamming}(n) = 0.54 + 0.46\cos(2\pi n/N) \quad (124)$$

$$\text{hann}(n) = 0.5 + 0.5\cos(2\pi n/N) \quad (125)$$

$$w(n) = a + (1 - a)\cos(2\pi n/N) \quad (126)$$



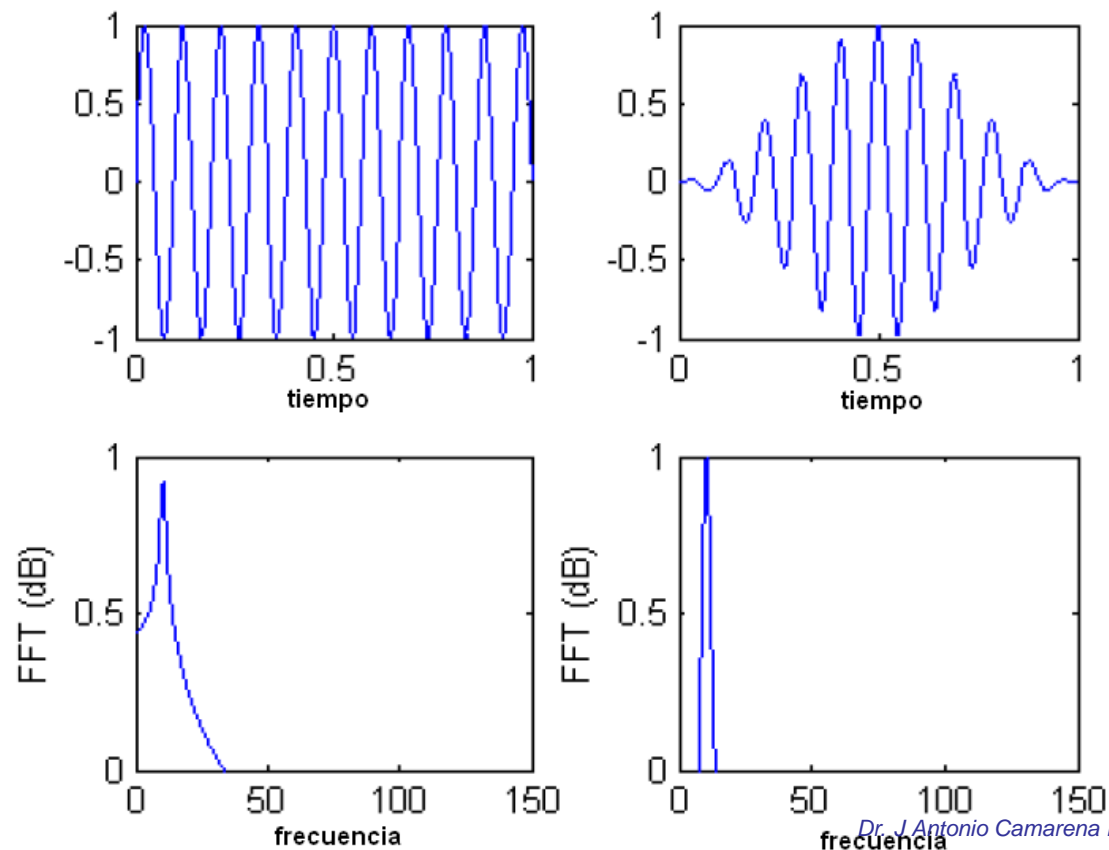


## El efecto "leakage"(escurrimiento)

- 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 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),

## El efecto "leakage"(escurrimiento)

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 y la transformada de Fourier de la misma senoide después de aplicarle la ventana de Hamming

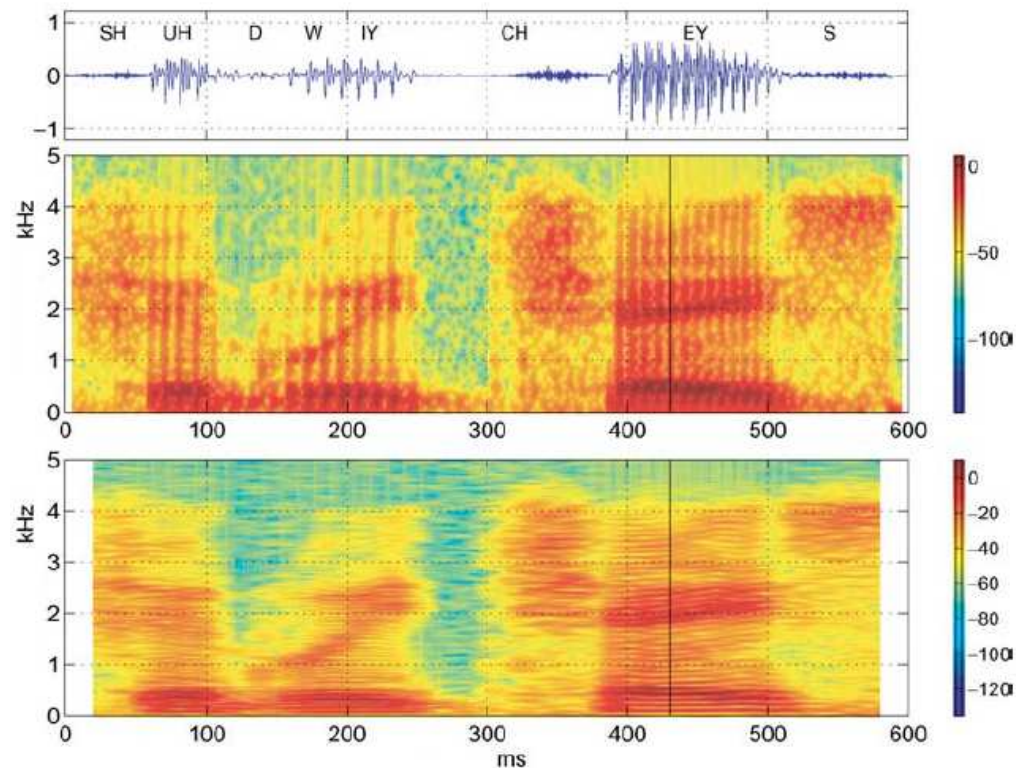


# Determinación del espectrograma mediante la Transformada de Fourier de tiempo corto

- Es necesario saber de que forma los contenidos de frecuencia de la señal de voz evolucionan en el tiempo.
- 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.
- Para apreciar de mejor manera el espectrograma se expresa la magnitud de los coeficientes en decibeles, es decir  $20\log_{10}(|X(k)|)$  en lugar de  $|X(k)|$ .

# Determinación del espectrograma mediante la Transformada de Fourier de tiempo corto

Espectrograma con marcos de 10ms y en la parte inferior se muestra el espectrograma con marcos de 40ms. A mayor resolución en el dominio del tiempo se tiene menor resolución en el dominio de la frecuencia y viceversa.





# Determinación del espectrograma de la señal de voz mediante Bancos de Filtros

- Un espectrograma con baja resolución en frecuencia 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.
- El espectrograma obtenido de esta manera presenta muchas ventajas, primero
  - Es fácil de implementar en hardware
  - Es paralelizable
  - 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 más deseable.

## 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.
- 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. Cada banda crítica tiene un ancho de un Bark,
- Para convertir Hertz a Barks podemos usar

$$z = 13 \tan^{-1} \left( \frac{0.76f}{1000} \right) + 3.5 \tan^{-1} \left( \frac{f}{750} \right)^2 \quad (127)$$

## Escala de Bark

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
:	:	:
24	15500	20000

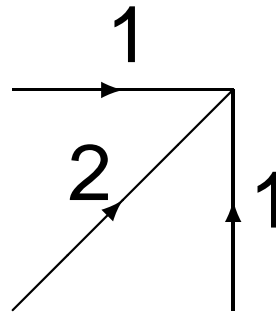
# Doblado Dinámico en Tiempo

- 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.
- Posiblemente 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  $(n - 1, m)$
- 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  $(n, m)$ ) tendrán todas el mismo costo de 2.
- Otras restricciones locales definidas por Itakura pueden usarse.



## Doblado Dinámico en Tiempo

Restricción local simétrica de primer orden



## Doblado Dinámico en Tiempo

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

$$D_{n,m} = \sum_{p=1}^n d_{R(p),T(w(p))} \quad (128)$$

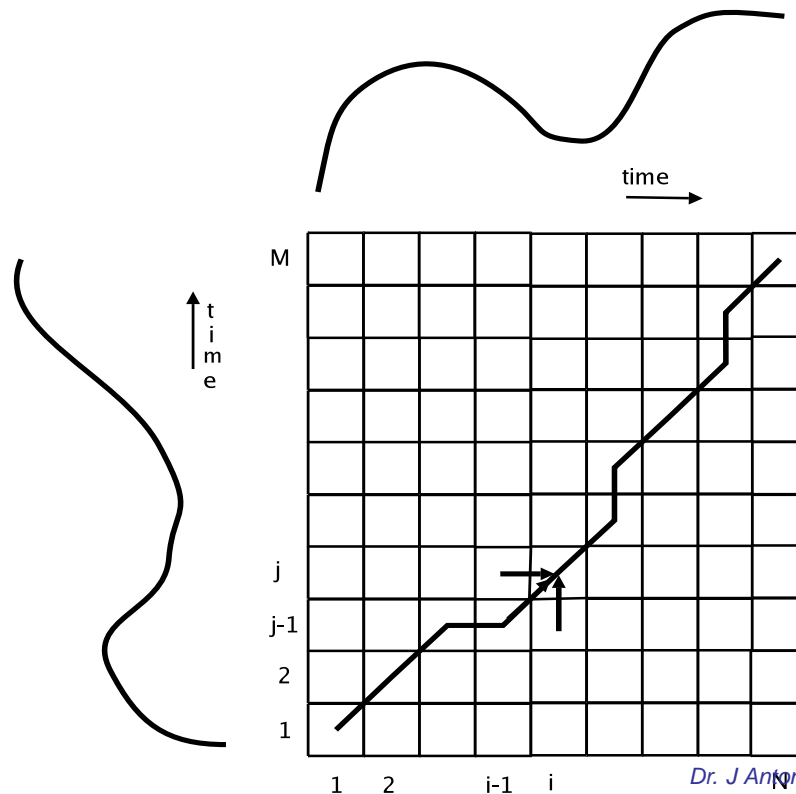
$D_{N,M}$  puede calcularse utilizando la recurrencia definida por las ecuaciones

$$\begin{aligned} D_{0,0} &= d_{0,0} \\ D_{i,0} &= d_{i,0} + D_{i-1,0} \\ D_{0,j} &= d_{0,j} + D_{0,j-1} \end{aligned}$$

$$D_{i,j} = \min \begin{cases} D_{i-1,j-1} + 2d_{i,j} \\ D_{i-1,j} + d_{i,j} \\ D_{i,j-1} + d_{i,j} \end{cases} \quad (129)$$

# Doblado Dinámico en Tiempo

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  $M$ . 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  $D_{N,M}/(N + M)$ .





## 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)



## Distancia de Manhattan

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 Manhattan recibe este nombre debido a que en una ciudad donde las calles son rectas (como en Manhattan) la distancia entre dos lugares de la misma ciudad la medimos en cuadras pues no podemos ir en línea recta.

$$d(a, b) = \sum_{i=1}^d |a_i - b_i| \quad (130)$$



## 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

$$d(a, b) = \sqrt{\sum_{i=1}^d (a_i - b_i)^2} \quad (131)$$



## Distancias LP

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

$$d(a, b) = \left( \sum_{i=1}^d (a_i - b_i)^p \right)^{1/p} \quad (132)$$

## 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 el ángulo menor que hay entre los dos vectores bajo comparación, el producto punto entre dos vectores se determina mediante:

$$a \cdot b = |a||b|\cos\theta \quad (133)$$

donde  $\theta$  es el ángulo menor que hay entre los vectores  $a$  y  $b$

$$\cos\theta = \frac{A \cdot B}{|a||b|} = \frac{a_1b_1 + a_2b_2 + \dots + a_db_d}{\sqrt{a_1^2 + a_2^2 + \dots + a_d^2}\sqrt{b_1^2 + b_2^2 + \dots + b_d^2}} \quad (134)$$



## Distancia coseno

$$\cos\theta = \frac{\sum_{i=1}^d a_i b_i}{\sqrt{\sum_{i=1}^d a_i^2} \sqrt{\sum_{i=1}^d b_i^2}} \quad (135)$$

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:

$$d(a, b) = 1 - |\cos\theta| = 1 - \left| \frac{\sum_{i=1}^d a_i b_i}{\sqrt{\sum_{i=1}^d a_i^2} \sqrt{\sum_{i=1}^d b_i^2}} \right| \quad (136)$$

## Criterio del vecino más cercano

- Para identificar un patrón al que llamaremos *consulta* se puede buscar de entre los patrones que conforman el diccionario 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 más 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 patrón más cercano comparando solo con algunos patrones del diccionario denominados *pivotes*.



## Criterio de los K-vecinos

- 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
- 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
- Se requiere que el diccionario tenga por cada clase mas de un patron lo cual complica la implementación, por ejemplo implica coleccionar por cada palabra del diccionario no solo una sino varias elocuciones, esto también incrementará el tamaño de la colección de patrones complicando las búsquedas.

# Implementación de un sistema de reconocimiento de palabras aisladas mediante espectrogramas

- Es posible identificar una palabra con solo ver su espectrograma
- Trasladar un problema de identificación de una señal de audio a un problema de reconocimiento de imágenes
- Una herramienta para los sordos (que no sean ciegos)
- Los espectrogramas se pueden utilizar como una caracterización de la señal de voz para realizar reconocimiento de palabras aisladas.
- 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 magnitudes 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.

# Implementación de un sistema de reconocimiento de palabras aisladas mediante espectrogramas

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:

$$E = \sum_{k=k_1}^{k_2} |X(k)|^2 \quad (137)$$

donde  $k_1$  es el índice del coeficiente espectral correspondiente al inicio de la banda y  $k_2$  es el índice del coeficiente espectral correspondiente al final de la banda en cuestión.

# Implementación de un sistema de reconocimiento de palabras aisladas mediante espectrogramas

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:

```
<Etiqueta1> <Num de marcos>
```

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

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

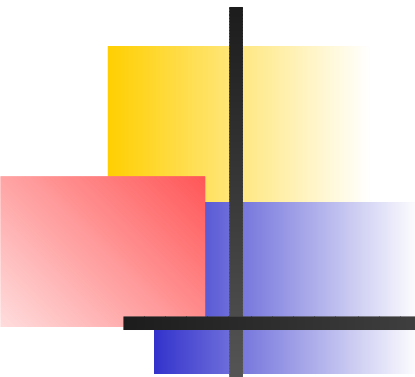
```
:
```

```
<Etiqueta2> <Num de marcos>
```

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

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

```
:
```



# Implementación de un sistema de reconocimiento de palabras aisladas mediante espectrogramas

- 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
- 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
- Detecta la entrada de diccionario con la menor distancia, si esta distancia es menor que cierto valor umbral, entonces reporta la ocurrencia de la palabra leyendo su etiqueta del diccionario
- Si la distancia es mayor al umbral declarar que la palabra es desconocida.

# 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 si tiene estructural fractal.
- La Transformada Continua de Fourier (CFT por sus siglas en inglés) no tiene semejante restricción,
- La CFT no se puede evaluar pues tiene un kernel infinito
- Una discretización de la CFT fue propuesta por el Dr. Campos



# Evaluación de la transformada continua de Fourier en los ceros de los polinomios de Hermite

1. Convertir la secuencia  $0,1,2,\dots,N-1$  a  $N$  valores equiespaciados desde  $-\pi$  hasta  $\pi$
2. Ajustar el segmento de la señal de voz a un polinomio trigonométrico de grado  $M < N/2$  para encontrar  $a_j$  and  $b_j$ .

$$\frac{a_0}{2} + \sum_{j=1}^M [a_j \cos(jx) + b_j \sin(jx)] \quad (138)$$

$$a_j = \frac{2}{N} \sum_{k=1}^N [f(x_k) \cos(jx_k)] \quad \forall \quad j = 0, 1, \dots, M \quad (139)$$

$$b_j = \frac{2}{N} \sum_{k=1}^N [f(x_k) \sin(jx_k)] \quad \forall \quad j = 1, 2, \dots, M \quad (140)$$

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

# Evaluación de la transformada continua de Fourier en los ceros de los polinomios de Hermite

4. Construya el Kernel de Fourier (matriz  $F$ ). 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.

$$F_{i,j} = \frac{\pi}{\sqrt{2n}} \sqrt[4]{\frac{4n+3-x_j^2}{4n+3-x_i^2}} [\cos(x_i x_j) + j \sin(x_i x_j)] \quad (141)$$

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.



# Procesamiento Homomórfico de la señal de VOZ

- 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.

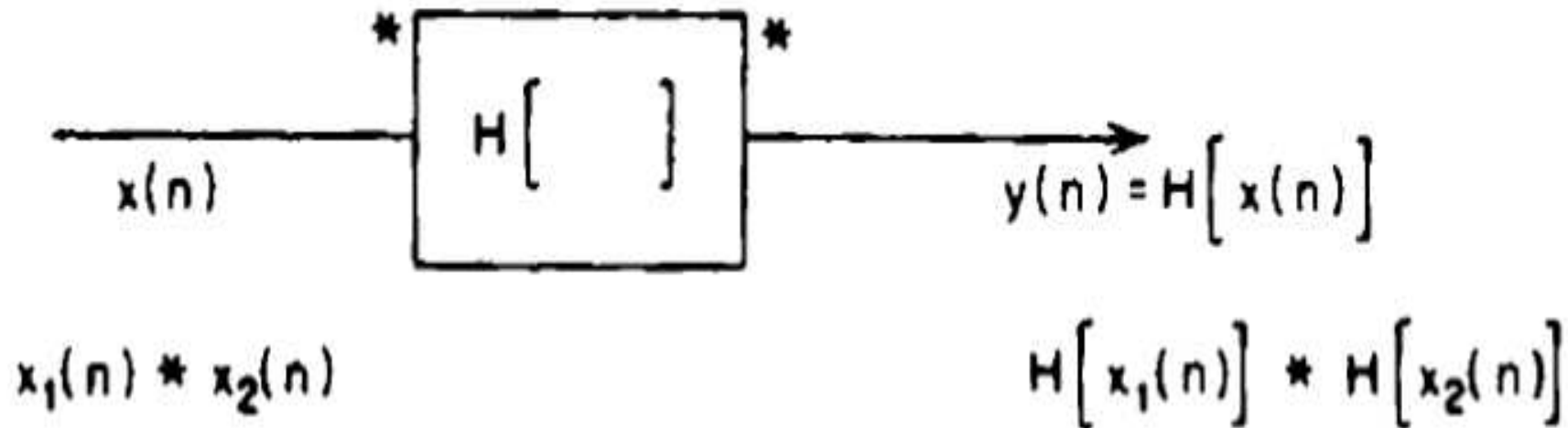


# Procesamiento Homomórfico de la señal de VOZ

- Un sistema homomórfico para la convolución obedece un principio generalizado de superposición (donde la operación suma es reemplazada por la operación convolución)
- 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 a la clase de sistemas que obedecen el principio generalizado de superposición como sistemas homomórficos para la convolución.

# Procesamiento Homomórfico de la señal de VOZ

Representación de un sistema homomórfico para la convolución.



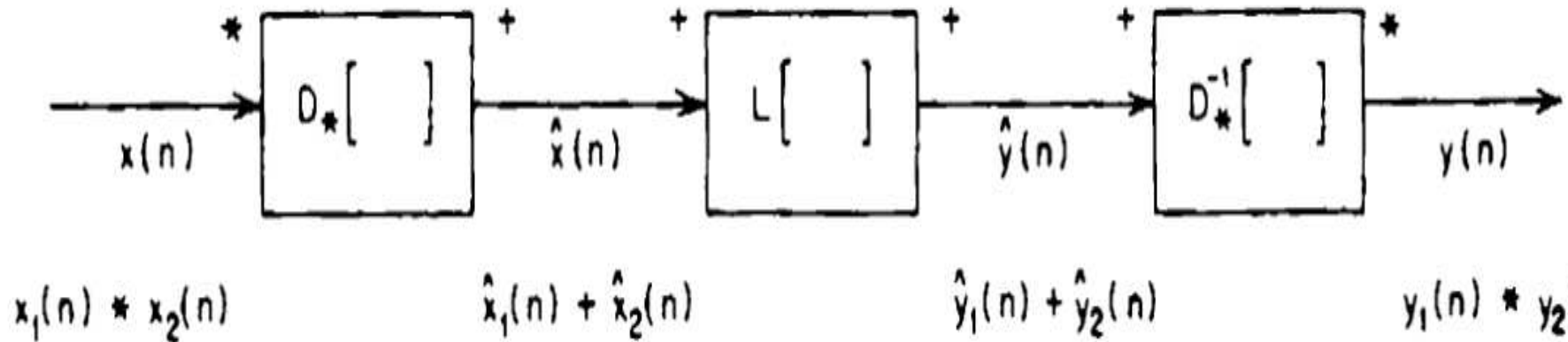


# Procesamiento Homomórfico de la señal de VOZ

- Todo sistema homomorfo 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 principio 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.

# Procesamiento Homomórfico de la señal de VOZ

Los tres sistemas, al sistema  $D_*[\ ]$  se le denomina *Sistema característico para deconvolución homomórfica*.



Forma canónica de un sistema para deconvolución homomórfica



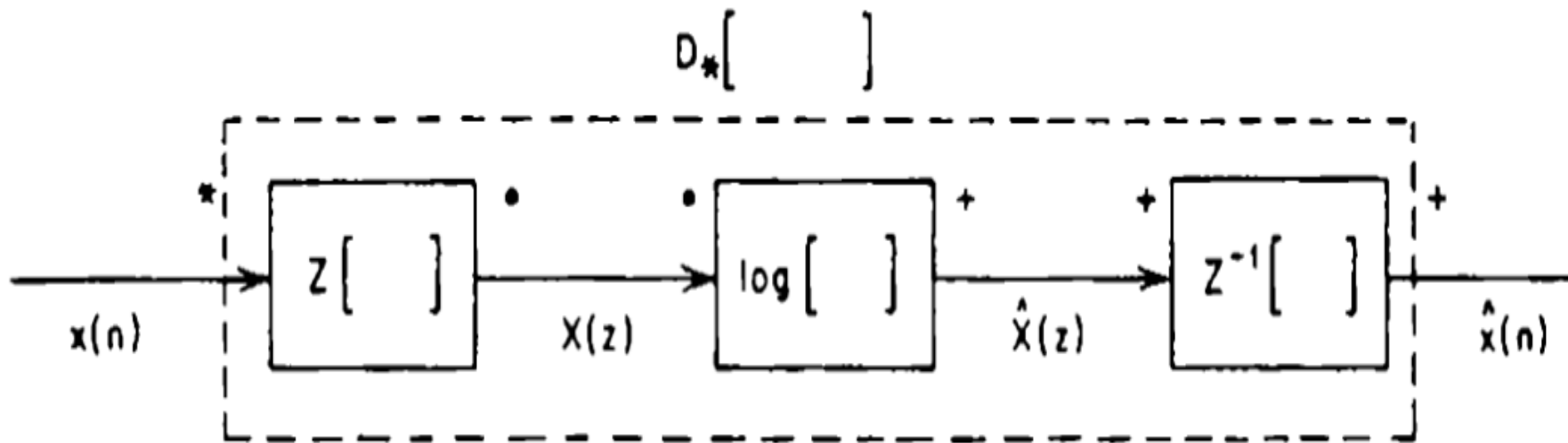
# Procesamiento Homomórfico de la señal de VOZ

- La Transformada  $Z$  de la salida de un sistema característico debe ser una combinación aditiva de transformadas  $Z$
- 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 basa en el hecho de que el logaritmo de un producto es igual a la suma de los logaritmos.



# Procesamiento Homomórfico de la señal de VOZ

Representación del sistema característico para deconvolución homomórfica



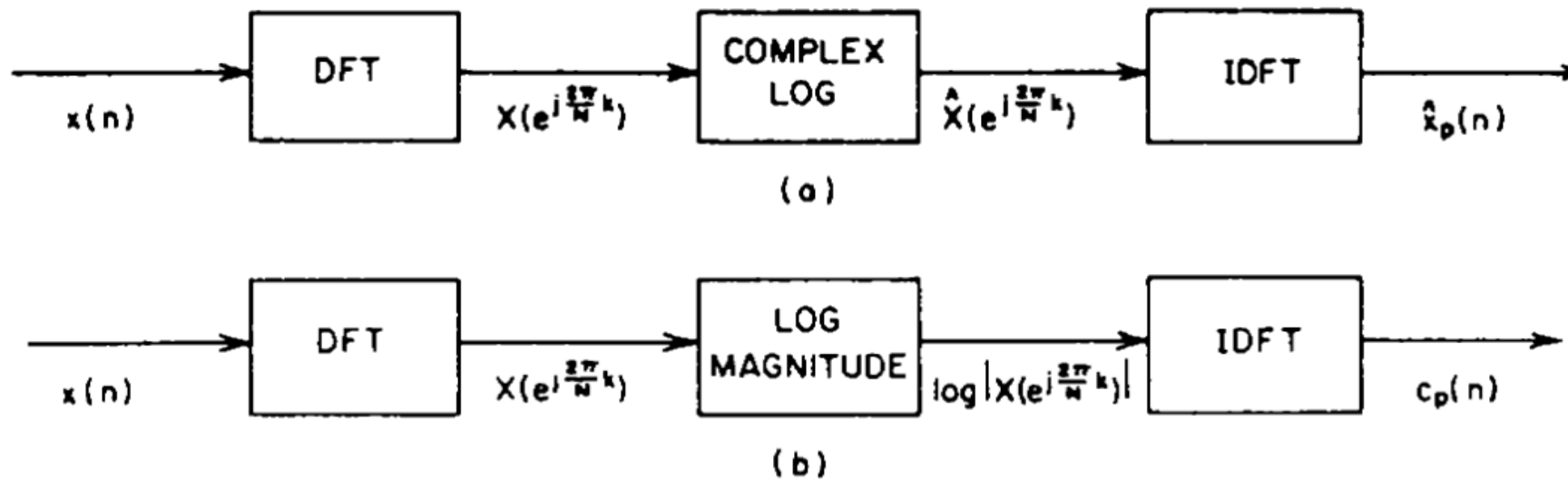


## El Cepstrum y el Cepstrum complejo

- 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.

## El Cepstrum y el Cepstrum complejo

Determinación del Cepstrum y del Cepstrum complejo.





## Aplicación a la estimación del tono

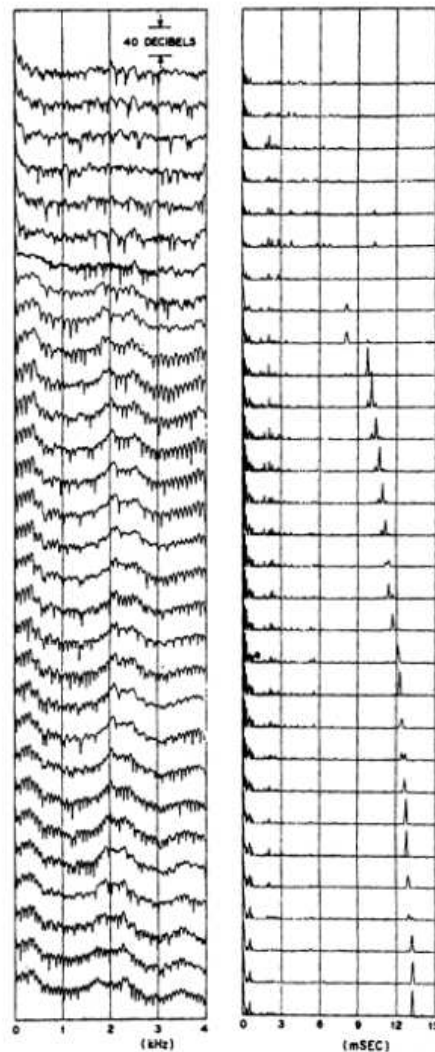
- El cepstrum permite distinguir entre sonidos vocalizados y no-vocalizados
- El tono se puede determinar a partir del cepstrum.
- Para sonidos vocalizados ocurre un pico en el cepstrum ubicado en una posición que coincide con el periodo de la frecuencia fundamental de la señal de voz
- Este pico no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



## Aplicación a la estimación del 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
- 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.

# Aplicación a la estimación del tono

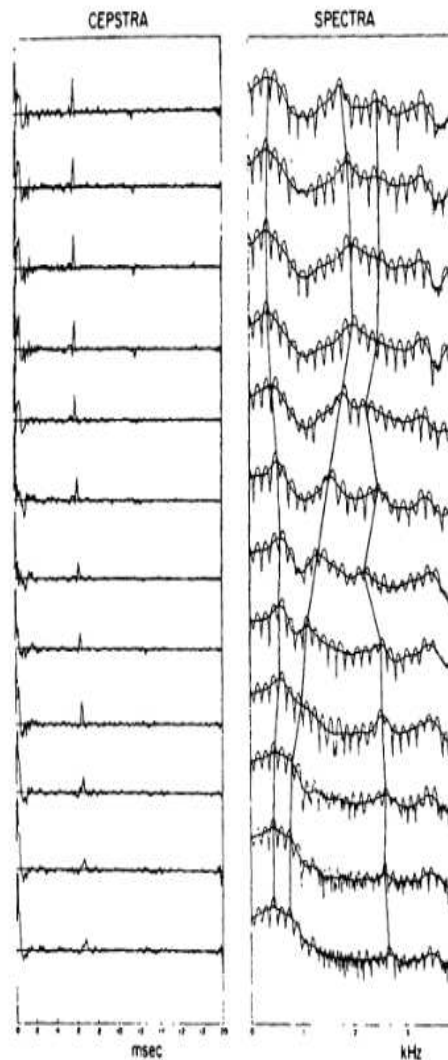




## Aplicación a la estimación de los formantes

- Podemos asumir que la parte inferior del cepstrum corresponde principalmente al tracto vocal, al pulso glotal y la radiación de la voz
- 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.
- 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.

# Aplicación a la estimación de los formantes





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

Los coeficientes cepstrales de Mel se determinan de acuerdo a ISP (Intelligent Sound Implementation). 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:

$$X(k) = \sum_{n=0}^{N-1} [x(n)w(n)]e^{-j2\pi kn/N} \quad k = 0, 1, \dots, N - 1 \quad (142)$$

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:

$$w(n) = 0.54 - 0.46\cos(\pi n/N) \quad (143)$$

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

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:

$$X'(m) = \ln \left( \sum_{k=0}^{N-1} [|X(k)| H(k, m)] \right) \quad (144)$$

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



## La escala de Mel

- La Escala Mel fué 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.
- Cuatro octavas en la escala de hertz 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.



## La escala de Mel

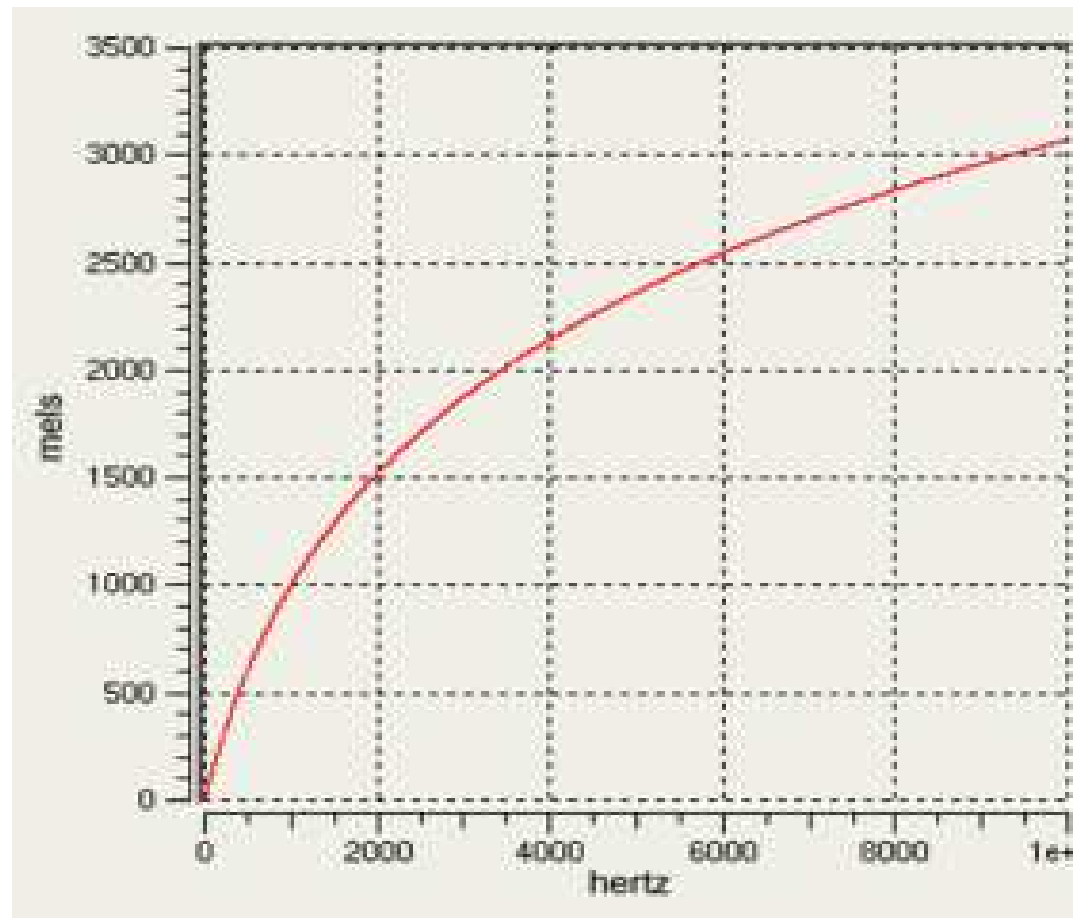
---

Para convertir  $f$  de Hertz a Mels se emplea:

$$\phi = 2595 \log_{10} \left( \frac{f}{700} + 1 \right) \quad (145)$$

que es una buena aproximación muy utilizada

# La escala de Mel



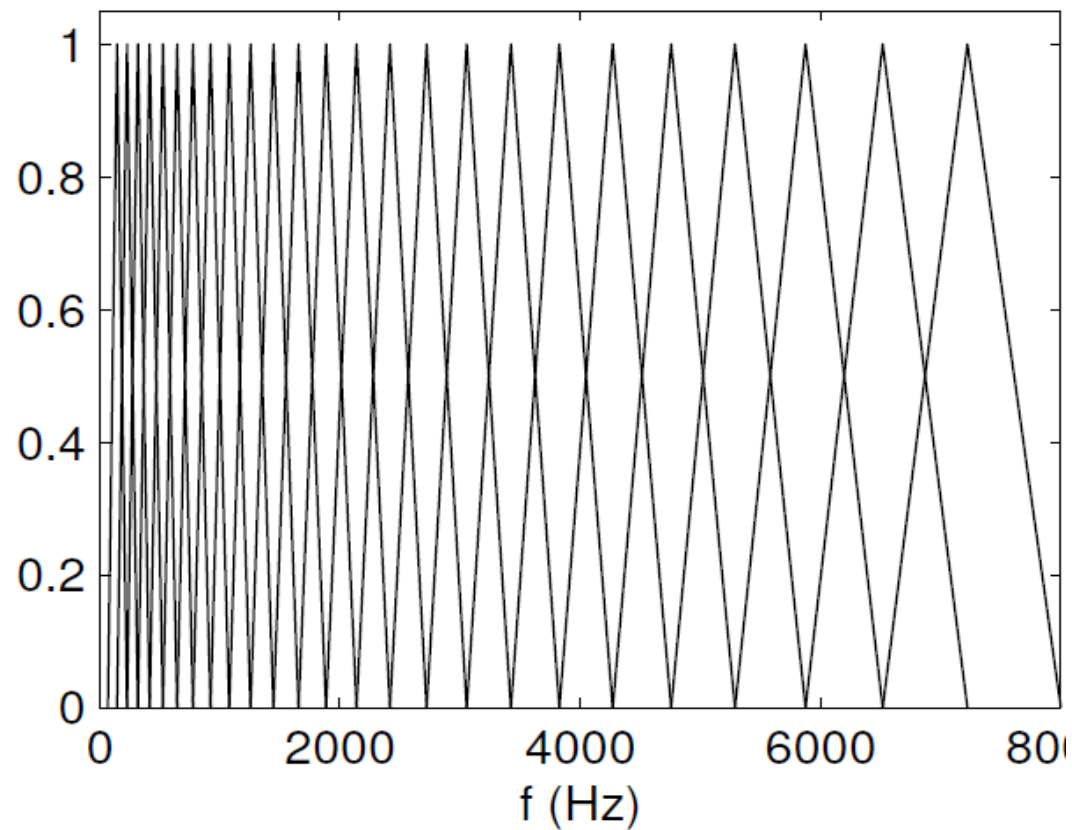
## Filtros triangulares de Mel

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

$$H(k, m) = \begin{cases} 0 & f(k) < f_c(m-1) \\ \frac{f(k) - f_c(m-1)}{f_c(m) - f_c(m-1)} & f_c(m-1) \leq f(k) < f_c(m) \\ \frac{f_c(m+1) - f(k)}{f_c(m+1) - f_c(m)} & f_c(m) \leq f(k) < f_c(m+1) \\ 0 & f(k) \geq f_c(m+1) \end{cases} \quad (146)$$

# Filtros triangulares de Mel

De acuerdo a las especificaciones del HTK (Hidden Markov Model Kit)



## Filtros triangulares de Mel

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:

$$\Delta\phi = \frac{\phi_{max} - \phi_{min}}{M + 1} \quad (147)$$

donde  $\phi_{max}$  y  $\phi_{min}$  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:

$$\phi_c(m) = m\Delta\phi \quad m = 1, 2, \dots, M \quad (148)$$

Para obtener las frecuencias centrales en Hertz utilizamos:

$$f_c(m) = 700 \left( 10^{\phi_c(m)/2595} - 1 \right) \quad (149)$$



## 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:

$$c(l) = \sum_{m=1}^M X'(m) \cos(l\pi(m - \frac{1}{2})/M) \quad l = 1, 2, \dots, M \quad (150)$$

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

# 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.
- 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 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.
- 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.

# 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

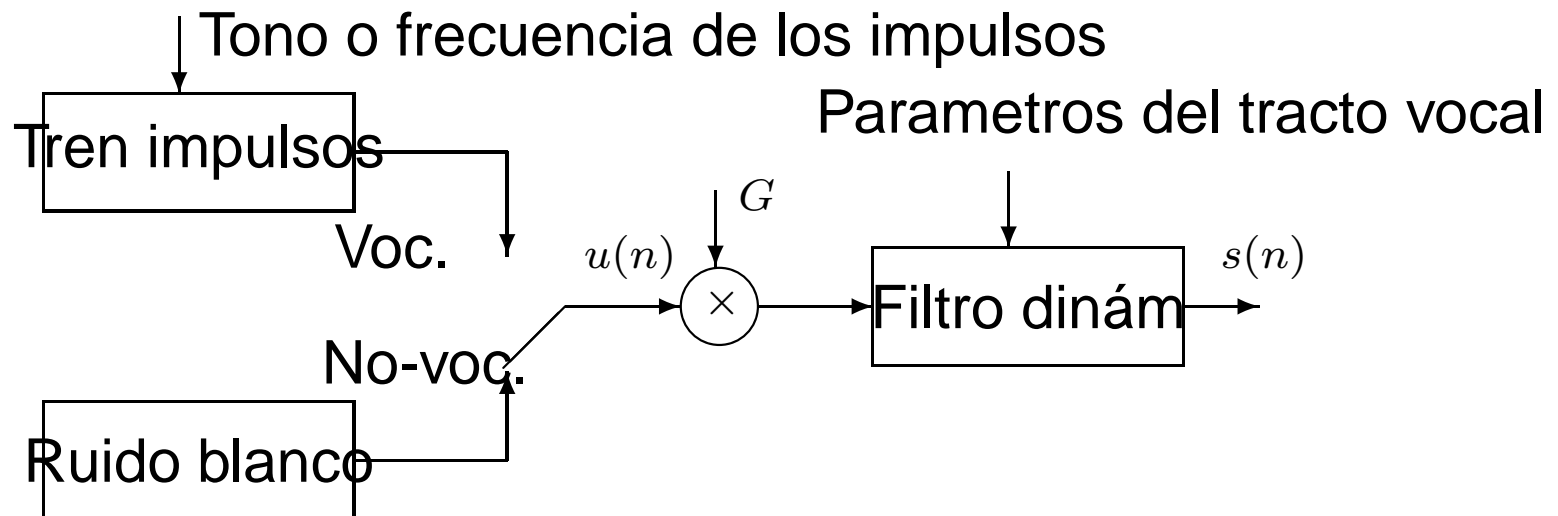


# Principios del análisis lineal predictivo

- 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 determinar los coeficientes predictores se minimiza la suma de los cuadrados de las diferencias entre las muestras obtenidas mediante 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.

# Principios del análisis lineal predictivo

$F$  frecuencia 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.



# Principios del análisis lineal predictivo

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

$$\frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^p a_k z^{-k}} \quad (151)$$

Aplicando transformada Z inversa:

$$s(n) = \sum_{k=1}^p a_k s(n - k) + Gu(n) \quad (152)$$

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

$$\tilde{s}(n) = \sum_{k=1}^p \alpha_k s(n - k) + Gu(n) \quad (153)$$

# Principios del análisis lineal predictivo

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

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^p \alpha_k s(n-k) + Gu(n) \quad (154)$$

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:

$$E_n = \sum_m e_n^2(m) = \sum_m [s_n(m) - \sum_{k=1}^p s_n(m-k)]^2 \quad (155)$$

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  $s(n+m)$ .

# Principios del análisis lineal predictivo

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

$$2 \sum_m \left[ S_n(m) - \sum_{k=1}^p \alpha_k s_n(m-k) \right] \frac{\partial \left[ - \sum_{k=1}^p \alpha_k s_n(m-k) \right]}{\partial \alpha_i} = 0 \quad i = 1, 2, \dots, p \quad (156)$$

pero:

$$\frac{\partial \left[ - \sum_{k=1}^p \alpha_k s_n(m-k) \right]}{\partial \alpha_i} = -s_n(m-i) \quad i = 1, 2, \dots, p \quad (157)$$

por lo cual:

$$2 \sum_m \left[ s_n(m) - \sum_{k=1}^p \alpha_k s_n(m-k) \right] [-s_n(m-i)] = 0 \quad i = 1, 2, \dots, p \quad (158)$$



# Principios del análisis lineal predictivo

de donde:

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^p \alpha_k \sum_m s_n(m-i)s_n(m-k) \quad i = 1, 2, \dots, p \quad (159)$$

Definiendo:

$$\phi_n(i, k) = \sum_m s_n(m-i)s_n(m-k) \quad (160)$$

Podemos escribirla en forma compacta como:

$$\sum_{k=1}^p \alpha_k \phi_n(i, k) = \phi_n(i, 0) \quad i = 1, 2, \dots, p \quad (161)$$

# 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  $0 \leq k \leq p$
- Luego resolver el sistema de ecuaciones.
- 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,
- 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

$$\phi_n(i, k) = \sum_{m=0}^{N+p-1} s_n(m-i)s_n(m-k) \quad i = 1, 2, \dots, p \quad k = 0, 1, 2, \dots, p \quad (162)$$

## Método de la Autocorrelación

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

$$\phi_n(i, k) = \sum_{r=0}^{N-1-(i-k)} s_n(r) s_n(r + i - k) \quad i = 1, 2, \dots, p \quad k = 0, 1, 2, \dots, p \quad (163)$$

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:

$$\phi_n(i, k) = R_n(|i - k|) \quad (164)$$

Entonces:

$$\sum_{k=1}^p \alpha_k R_n(|i - k|) = R_n(i) \quad i = 1, 2, \dots, p \quad (165)$$

## Solución de las ecuaciones LPC

Para encontrar los coeficientes LPC, debemos resolver el sistema de ecuaciones expresado en forma desarrollada:

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \dots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \dots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \dots & R_n(p-3) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ R_n(p) & R_n(p-1) & R_n(p-2) & \dots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \vdots \\ R_n(p) \end{bmatrix} \quad (166)$$

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

# Método recursivo de Durbin para solucionar las ecuaciones de autocorrelación

- La matriz de autocorrelación  $T$  es una matriz de Toeplitz 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*
- Este procedimiento comienza con un predictor de primer orden, es decir, de un solo coeficiente e incrementa el orden recursivamente utilizando las soluciones de orden inferior para obtener soluciones de órdenes superiores. La notación  $\alpha_i^p$  denota al  $i$ -ésimo coeficiente de un predictor de orden  $p$ .

# Método recursivo de Durbin para solucionar las ecuaciones de autocorrelación

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

$$\alpha_1^{(1)} R_n(0) = R_n(1) \quad (167)$$

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

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

$$E_n = R_n(0) - \sum_{k=1}^p \alpha_k R_n(k) \quad (168)$$

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

$$E_n^{(1)} = R_n(0) - \alpha_1^{(1)} R_n(1) \quad (169)$$

# Método recursivo de Durbin para solucionar las ecuaciones de autocorrelación

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

$$\begin{bmatrix} R_n(0) & R_n(1) \\ R_n(1) & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1^{(2)} \\ \alpha_2^{(2)} \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \end{bmatrix} \quad (170)$$

# Método recursivo de Durbin para solucionar las ecuaciones de autocorrelación

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

$$\alpha_2^{(2)} = \frac{R_n(0)R_n(2) - R_n(1)^2}{R_n^2(0) - R_n^2(1)} \quad (171)$$

Dividiendo numerador y denominador por  $R_n(0)$

$$\alpha_2^{(2)} = \frac{R_n(2) - R_n(1)^2/R_n(0)}{R_n(0) - R_n^2(1)/R_n(0)} \quad (172)$$

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

$$\alpha_2^{(2)} = \frac{R_n(2) - R_n(1)\alpha_1^{(1)}}{R_n(0) - R_n(1)\alpha_1^{(1)}} \quad (173)$$

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

$$\alpha_2^{(2)} = \frac{R_n(2) - R_n(1)\alpha_1^{(1)}}{E_n^{(1)}} \quad (174)$$



# Método recursivo de Durbin para solucionar las ecuaciones de autocorrelación

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  $p$

$$k_i = \left[ R_n(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R_n(i-j) \right] / E^{(i-1)}$$

$$\alpha_i^{(i)} = k_i$$

$$\text{Desde } j = 1 \text{ hasta } i - 1 \quad \alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}$$

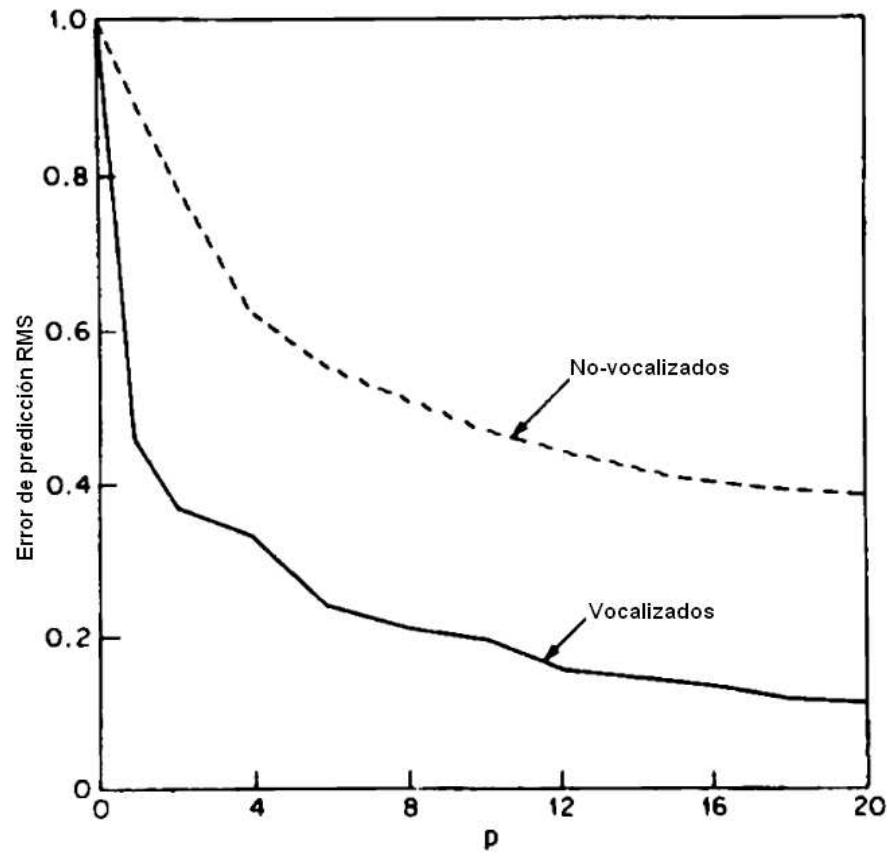
$$\text{Si } i < p \text{ entonces } E^{(i)} = (1 - k_i^2) E^{(i-1)}$$

# Método recursivo de Durbin para solucionar las ecuaciones de autocorrelación

- 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  $p$ .
- Ejemplo, para una frecuencia de muestreo de 8 KHz se recomienda  $p = 12$ .

# Método recursivo de Durbin para solucionar las ecuaciones de autocorrelación

Error de predicción vs  $p$



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

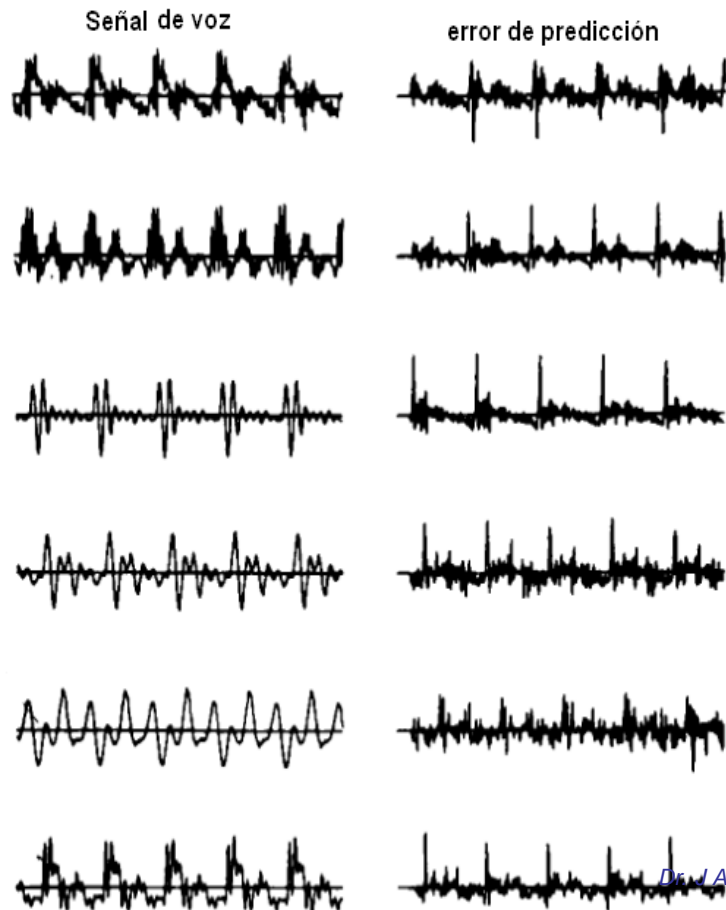
Como un producto colateral, el análisis LPC genera la señal de error  $e(n)$  definida como:

$$e(n) = s(n) - \sum_{k=1}^p \alpha_k s(n - k) = Gu(n) \quad (175)$$

- 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  $p$ .
- Es de esperarse entonces que el error de predicción aumente drásticamente cada vez que inicia otro periodo del tono para sonidos vocalizados
- 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 señal de error y su relación con la estimación del pulso glotal y aplicación a identificación de individuos por su voz

## La señal de error



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

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:

$$V_n = \frac{\sum_{m=0}^{N+p-1} e_n^2(m)}{\sum_{m=0}^{N-1} s_n^2(m)} \quad (176)$$

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

$$e(n) = - \sum_{k=0}^p \alpha_k s(n-k) \quad (177)$$

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

$$V_n = \sum_{i=0}^p \sum_{j=0}^p \alpha_i \frac{\phi_n(i, j)}{\phi_n(0, 0)} \alpha_j \quad (178)$$

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

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:

$$V_n = \prod_{i=1}^p (1 - k_i^2) \quad (179)$$

- 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.

# Interpretación en el dominio de la frecuencia del análisis lineal predictivo

Dado un conjunto de coeficientes podemos encontrar la respuesta a la frecuencia del modelo de producción de voz simplemente evaluando  $H(z)$  para  $z = e^{j\omega}$ ,

$$H(e^{j\omega}) = \frac{G}{1 - \sum_{k=1}^p \alpha_k e^{j\omega k}} = \frac{G}{A(e^{j\omega})} \quad (180)$$

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.



# Interpretación en el dominio de la frecuencia del error de predicción

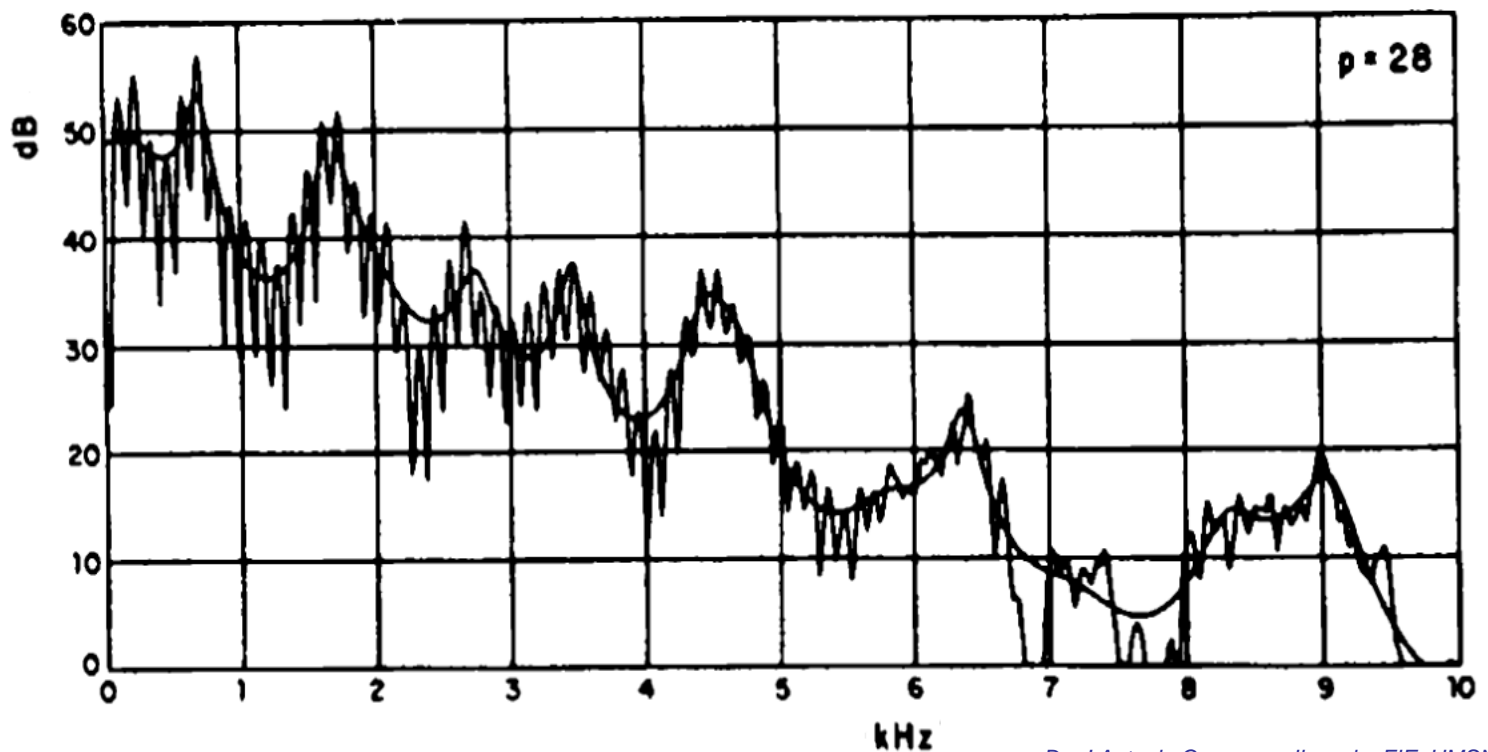
Para determinar los coeficientes de un predictor lineal, minimizamos el error de predicción de tiempo corto. El error de predicción puede hacerse arbitrariamente pequeño si incrementamos el orden del predictor lo suficiente. Es fácil mostrar que:

$$\lim_{p \rightarrow \infty} |H(e^{j\omega})|^2 = |X(e^{j\omega})|^2 \quad (181)$$

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,

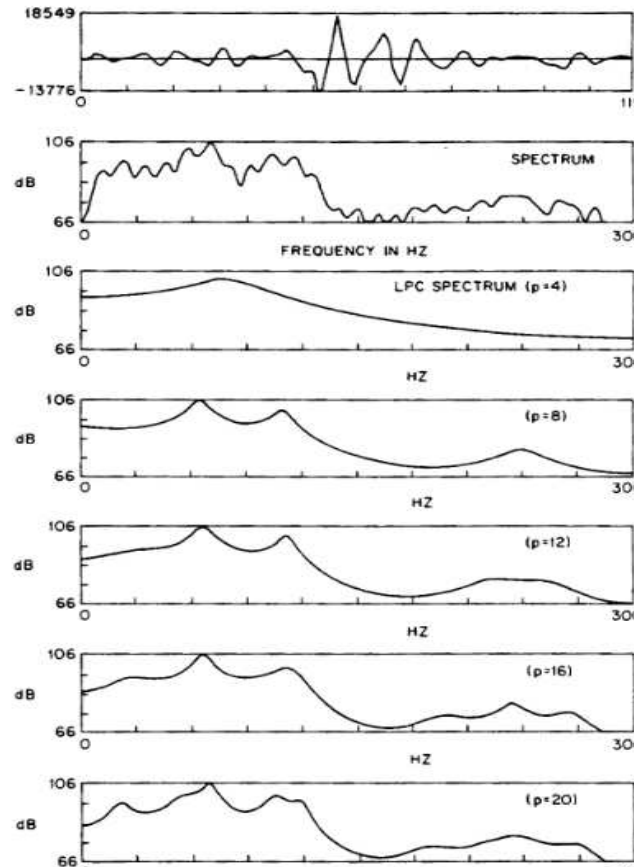
# Interpretación en el dominio de la frecuencia del error de predicción

Espectro  $20\log_{10}|X(e^{j\omega})|$  obtenido al realizar un análisis basado en la Transformada de Fourier. En la misma figura se muestra el espectro de  $20\log_{10}|H(e^{j\omega})|$  de la misma señal de voz para un predictor de orden  $p = 28$ .



# Interpretación en el dominio de la frecuencia del error de predicción

El orden  $p$  del predictor puede efectivamente controlar el grado de suavidad del espectro de la señal



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

Es posible modelar la producción de la voz mediante la concatenación de  $N$  tubos sin pérdidas. 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:

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \quad (182)$$

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 que la función sistema del modelo reticular es:

$$V(z) = \frac{\prod_{k=1}^N (1 + r_k) z^{-N/2}}{D(z)} \quad (183)$$

donde  $D(z)$  satisface la recursión:

$$D_0(z) = 1 \quad (184)$$

$$D_k(z) = D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1}) \quad (185)$$

$$D(z) = D_N(z) \quad (186)$$

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

Por otro lado, el polinomio

$$A(z) = 1 - \sum_{k=1}^p \alpha_k z^{-k} \quad (187)$$

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

$$A^0(z) = 1 \quad (188)$$

$$A^i(z) = A^{i-1}(z) - k_i z^{-i} A^{i-1}(z^{-1}) \quad (189)$$

$$A(z) = A^p(z) \quad (190)$$

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

$$A(z) = \frac{G}{A(z)} \quad (191)$$

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

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:

$$r_i = -k_i \quad (192)$$

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

$$A_{i+1} = A_i \frac{1 - k_i}{1 + k_i} \quad (193)$$



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

- Observe que los coeficientes PARCOR proporcionan una razón entre áreas de secciones adyacentes
- 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.
- 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.

## Los coeficientes PARCOR

A los valores  $k_i$  que se determinan de manera colateral por el Algoritmo 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 los métodos basados en retículas para determinar los coeficientes predictores.

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

$$-1 < k_i < 1 \quad \forall 1 \leq i \leq p \quad (194)$$



## Los coeficientes PARCOR

- 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 PARCOR

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

$$a_i^{(i)} = k_{(i)} \quad (195)$$

$$a_j^{(i)} = k_i a_{i-j}^{(i-1)} \quad 1 \leq j \leq i - 1 \quad (196)$$

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

$$\alpha_j = a_j^{(p)} \quad 1 \leq j \leq p \quad (197)$$

## Los coeficientes PARCOR

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

$$k_{(i)} = a_i^{(i)} \quad (198)$$

$$a_j^{(i-1)} = \frac{a_j^{(i)} + a_i^{(i)} a_{i-j}^{(i)}}{1 - k_i^2} \quad 1 \leq j \leq i - 1 \quad (199)$$

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

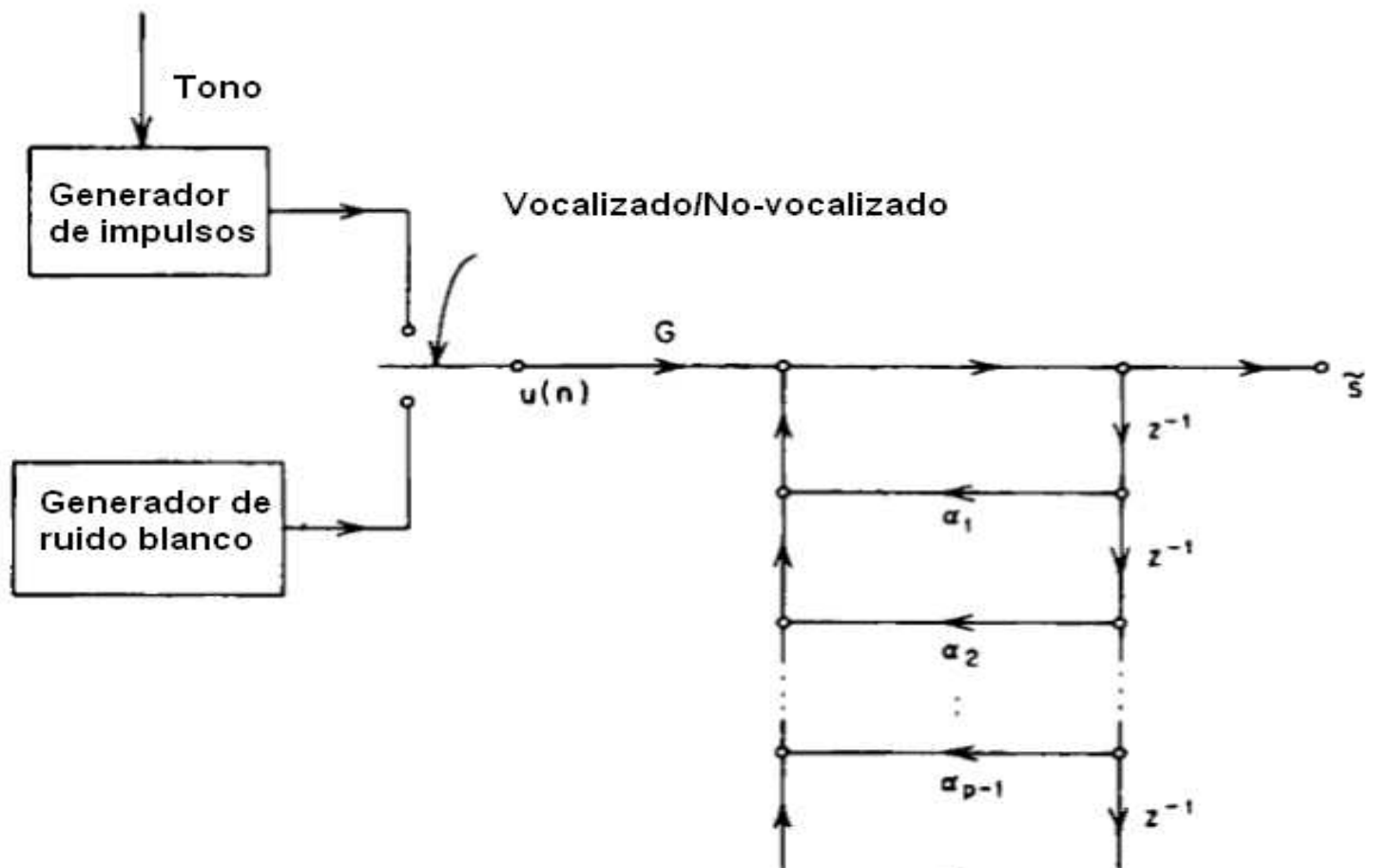
$$a_j^{(p)} = \alpha_j \quad 1 \leq j \leq p \quad (200)$$

# 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.
- Los parámetros de control variables en el tiempo que se requieren para el sintetizador son el tono, el booleano vocalizado/no-vocalizado, la ganancia 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:

$$\tilde{s}(n) = \sum_{k=1}^p \tilde{s}(n - k) + Gu(n) \quad (201)$$

# Síntesis de voz mediante parámetros LPC



# Síntesis de voz mediante parámetros LPC

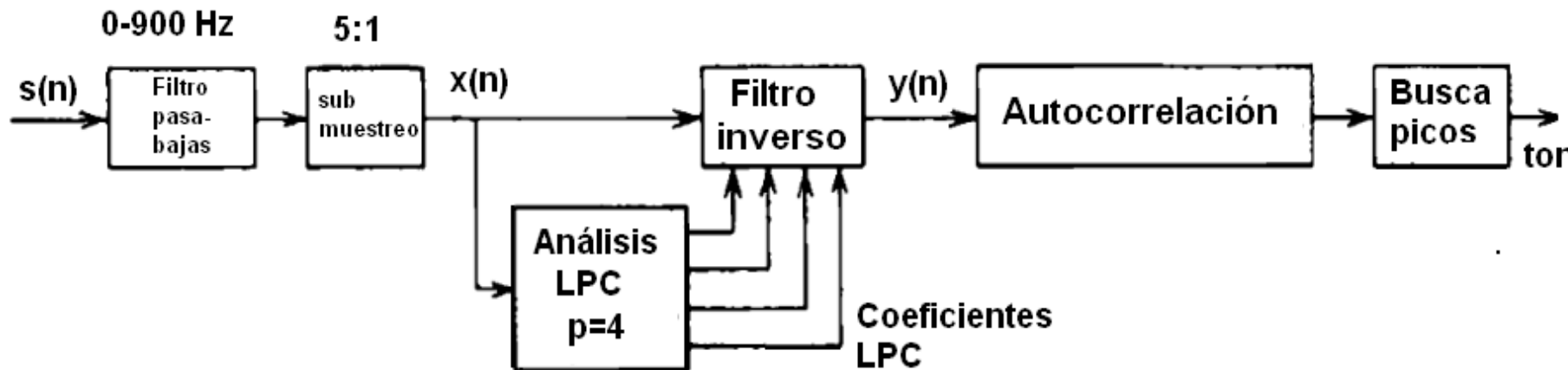
- Para sonidos no-vocalizados los parámetros simplemente se cambian en cada marco
- Para sonidos vocalizados, los parámetros son cambiados al inicio de cada periodo, a esto se le conoce como síntesis sincronizada al tono
- Se ha descubierto que es una síntesis mas efectiva que cuando se cambian los parámetros 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)
- Debido a restricciones de estabilidad, los coeficientes predictores no pueden interpolarse directamente.
- La razón es que la interpolación de dos conjuntos de coeficientes correspondientes a dos predictores estables puede producir un conjunto de coeficientes de un predictor inestable.
- La principal ventaja del sintetizador 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

# Síntesis de voz mediante parámetros LPC

- 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.

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



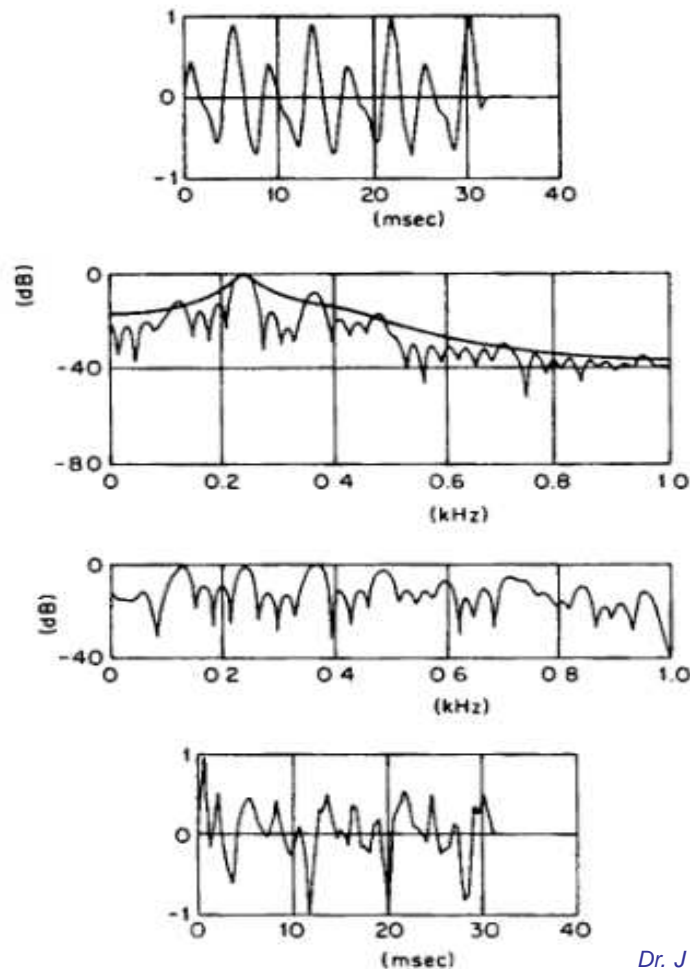


# Determinación del tono usando los coeficientes LPC

- 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.
- El propósito del análisis LPC es aplanar el espectro, enseguida se obtiene la función de autocorrelación de  $y(n)$  t' de donde se localiza el pico mas grande, cuya ubicación es la estimación del tono.

# Determinación del tono usando los coeficientes LPC

Señales involucradas en el Algoritmo SIFT para estimación del tono





# Determinación del tono usando los coeficientes LPC

- a. Muestra la forma de onda analizada
- b. Muestra su espectro en donde se aprecia que hay solo un formante en el rango de 250 Hz.
- 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.
- d. Muestra la señal sintetizada en el dominio del tiempo.
- e. Muestra la función de autocorrelación de la señal sintetizada de donde un tono de 8 ms aproximadamente se puede ver con claridad.

# 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
- Se sabe que el espectro de sonidos nasales no solo tiene polos sino ceros.

# 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.
- 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 cada columna corresponde con una banda crítica de Bark, mientras que en el reconocedor hecho en base a coeficientes LPC, cada columna corresponde con los coeficientes predictores.
- Conviene utilizar la distancia de Itakura, la cual fue diseñada específicamente para comparar vectores cuyos componentes son precisamente coeficientes LPC

# Reconocedor de palabras aisladas basado en coeficientes LPC

- La distancia de Itakura 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
- 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 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  $p = 12$ .
- Un reconocedor hecho en base a coeficientes LPC no requiere de aplicación de ventanas (aunque tampoco le perjudica)
- Como no se aplicará la FFT no es necesario usar marcos cuyo tamaño en muestras sea una potencia de 2, de todos modos conviene usar marcos de 30 ms con traslape de dos tercios de manera que los marcos avancen 10 ms cada vez.

## Distancia de Itakura

Hemos visto como tanto para reconocimiento de voz como para identificación de individuos por su voz es necesario comparar cuantitativa y eficientemente 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  $a = (1, \alpha_1, \alpha_2, \dots, \alpha_p)$  es el vector con los coeficientes LPC de un marco de voz y  $\hat{a} = (1, \hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_p)$ . Como  $D(a, \hat{a})$  es una medida de distancia se requiere que

$$D(a, \hat{a}) \geq 0 \quad (202)$$

y que:

$$D(a, a) = 0 \quad (203)$$

## Distancia de Itakura

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  $\Sigma$  la cual se determina mediante:

$$\Sigma = \frac{R^{-1}}{N} \hat{a} R \hat{a}^t \quad (204)$$

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.



## Distancia de Itakura

La probabilidad de obtener  $\hat{a}$  es:

$$P(\hat{a}|a) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-0.5(\hat{a}-a)\Sigma^{-1}(\hat{a}-a)^t} \quad (205)$$

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

$$D(\hat{a}, a) = (\hat{a} - a) \left[ N \frac{R}{\hat{a} R a^t} \right] (\hat{a} - a)^t \quad (206)$$

## Distancia de Itakura

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

$$D'(\hat{a}, a) = \log \left[ \frac{aRa^t}{\hat{a}R\hat{a}^t} \right] \quad (207)$$

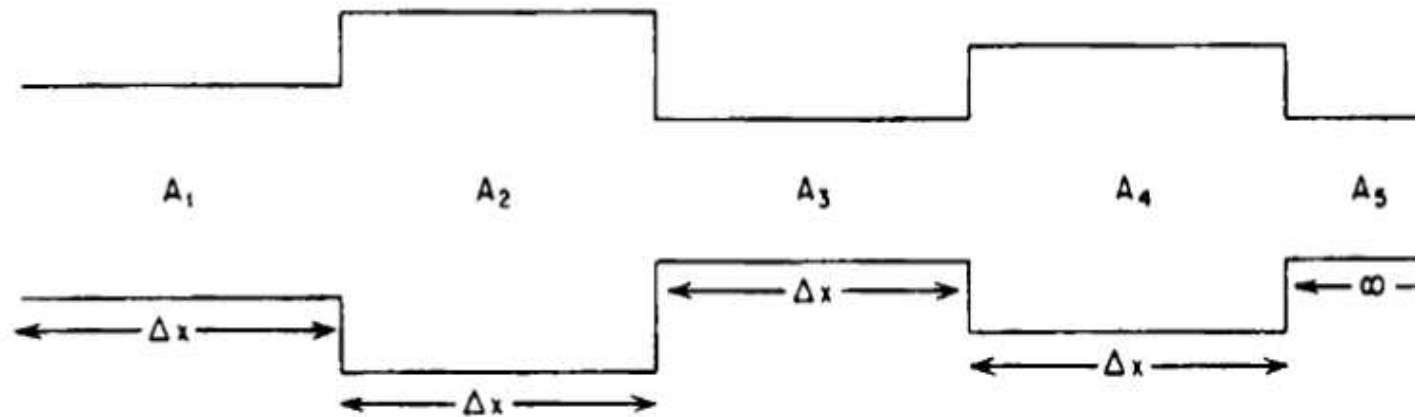
La clave para entender esta distancia es que dos vectores de coeficientes LPC deben tener una distancia corta entre ellos si fueron extraídos de segmentos de voz que suenan muy parecido

# Síntesis de voz basado en coeficientes LPC

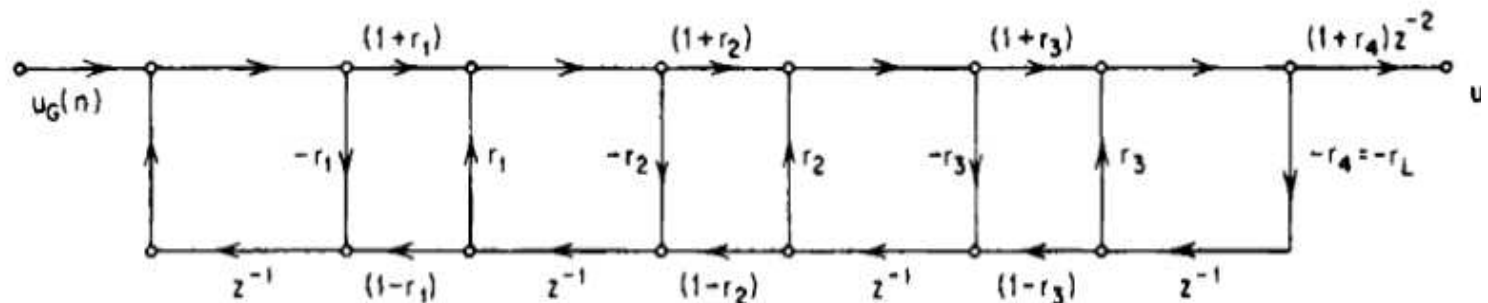
- La alternativa a la síntesis basada en coeficientes predictores que resulta mas atractiva es la síntesis basada en coeficientes de reflexión del modelo basado en concatenación de tubos sin pérdidas
- Se puede cambiar del sintetizador mostrado en la la parte que recibe  $u(n)$  y entrega  $\tilde{s}(n)$  por la red mostrada en la siguiente lámina
- La ventaja principal es que los multiplicadores son los coeficientes de reflexión  $r_i = -k_i$ , los cuales tienen la propiedad de estar acotados ( $|k_i| < 1$ )
- 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.

# Síntesis de voz basado en coeficientes LPC

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

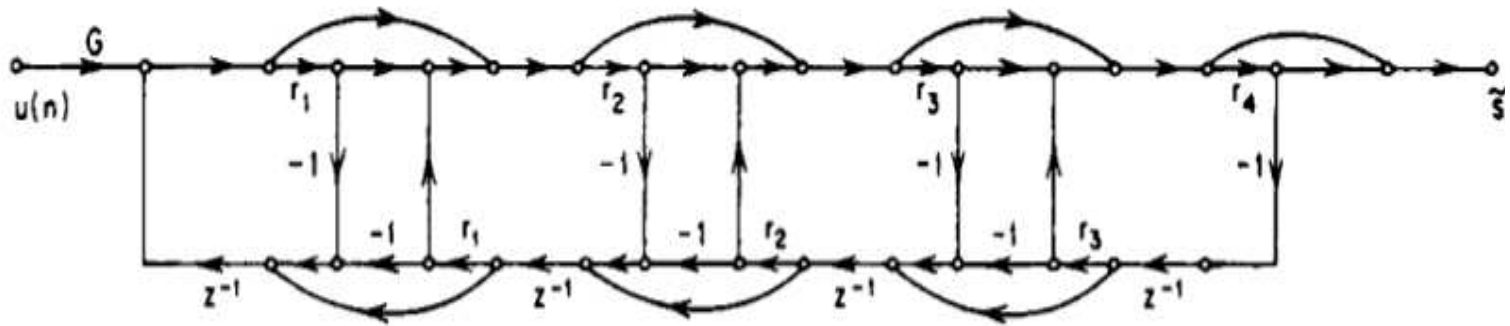


(a)

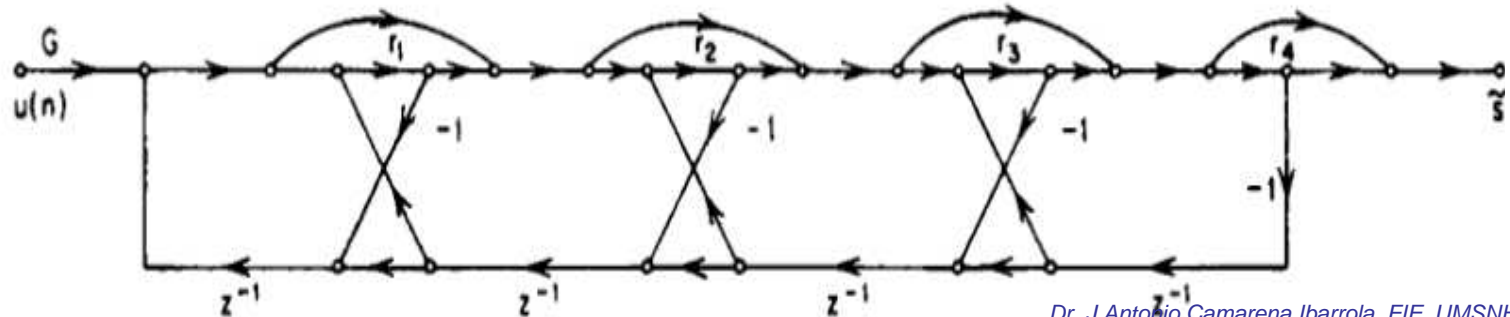


# Síntesis de voz basado en coeficientes LPC

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



(a)



# Síntesis de voz basado en coeficientes LPC

- Implementar el primer sintetizador de orden  $p$  basado en coeficientes reflectores requiere  $4p + 2$  multiplicaciones y  $2(p + 1)$  adiciones por muestra
- Es una desventaja respecto al sintetizador basado en coeficientes predictores del mismo orden de la que solo requiere  $p$  multiplicaciones y  $p$  adiciones
- El sintetizador de la Figura (a) que requiere  $2p - 1$  multiplicaciones y  $4p - 1$  adiciones
- El sintetizador de la Figura (b) que solo requiere  $p$  multiplicaciones y  $3p - 2$  adiciones.