

Análisis de la voz

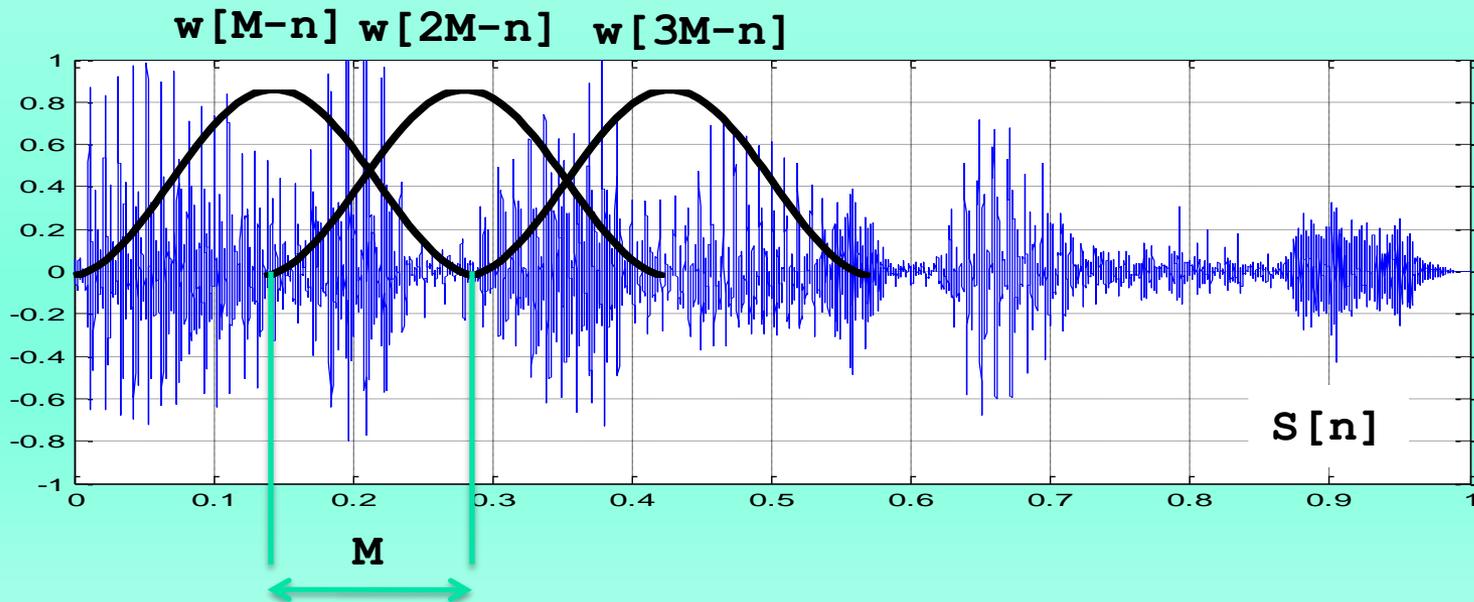
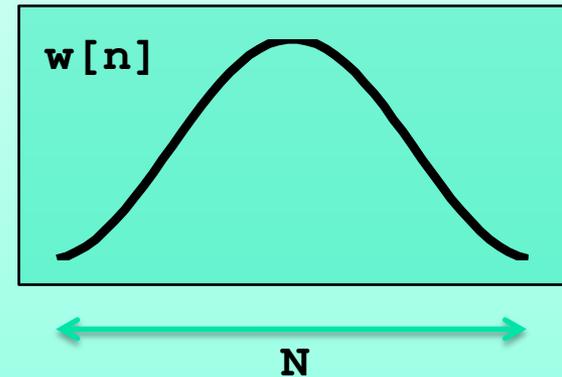
Parametrización

Análisis localizado de la voz

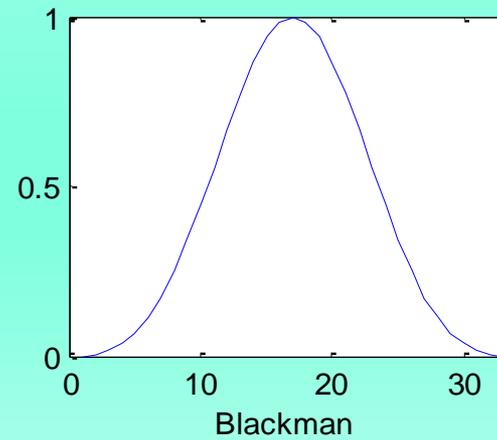
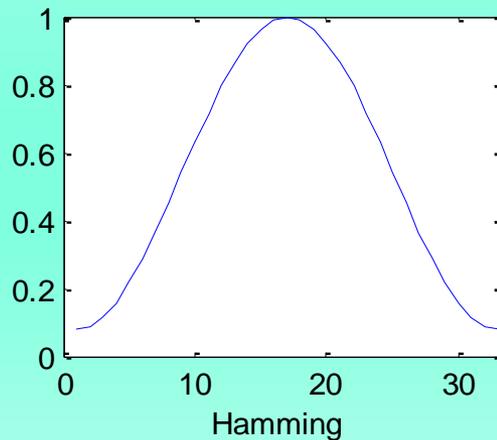
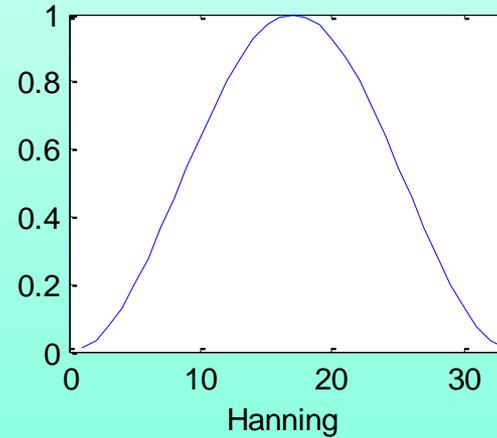
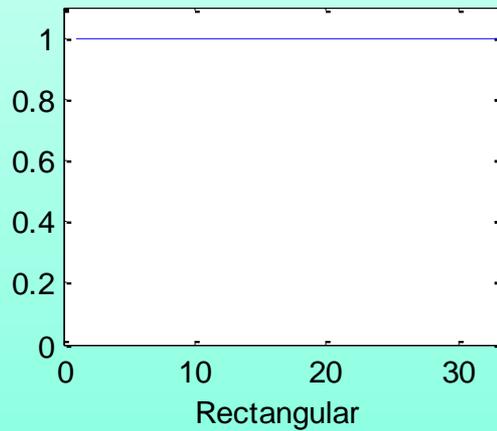
- La señal de voz solo presenta características pseudo-estacionarias a corto plazo
- Será necesario procesar la señal de voz en segmentos de corta duración: **Análisis Localizado**
- El mecanismo que nos permite realizar este análisis es el enventanado de la señal

Enventanado de la señal

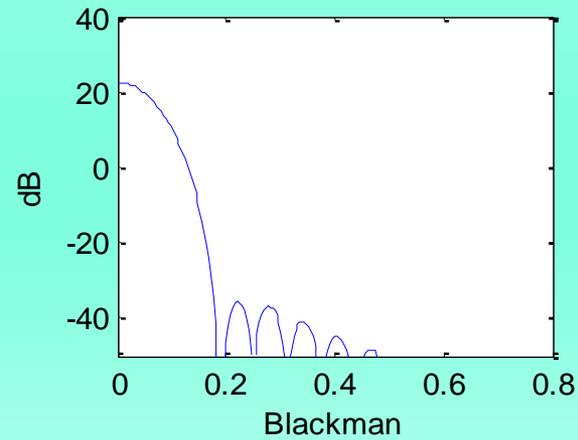
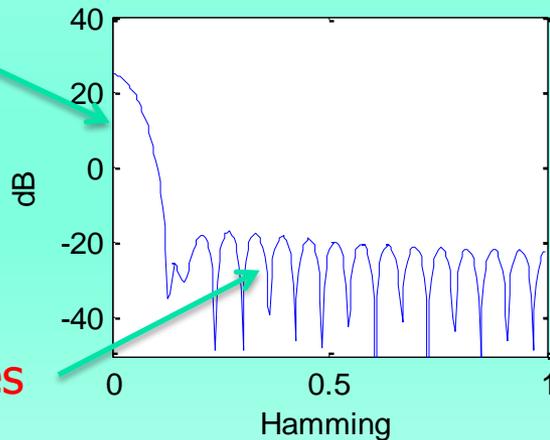
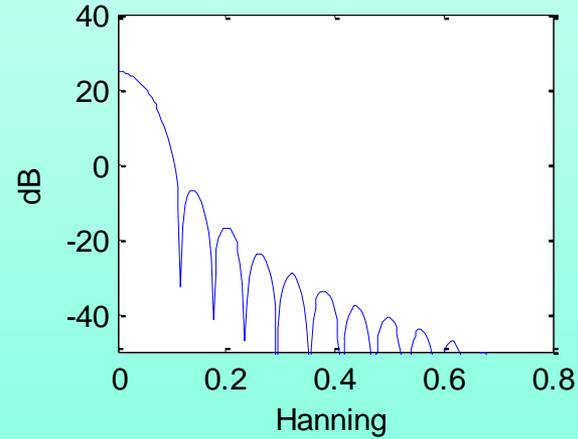
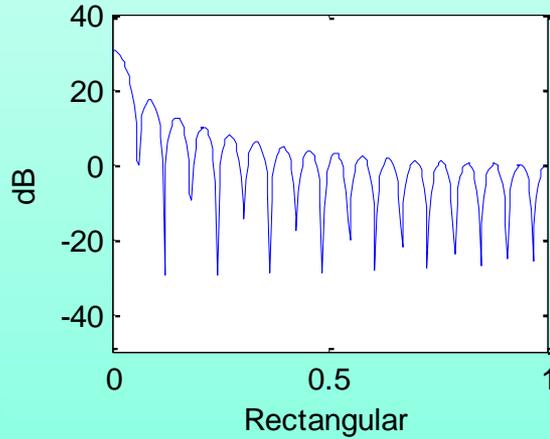
- $s[n]$: Señal de voz
- $w[n]$: Ventana de análisis
 - N : Tamaño de la ventana
 - M : Desplazamiento



- Profiles: Rectangular, Hanning, Hamming, Blackman...

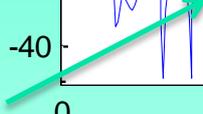


- Espectros de los perfiles:



Lóbulo principal

Lóbulos laterales



- Problemas del enventanado:
 - Produce derrame espectral (*leakage*).
 - El lóbulo principal dificulta la identificación de frecuencias cercanas entre sí.
 - Los lóbulos laterales introducen señal en frecuencias donde no debería haber nada.
- Se debe llegar a un compromiso entre el ancho del lóbulo principal y la minimización de los laterales.
- Generalmente se prefiere minimizar los lóbulos laterales.
- Perfiles típicos para voz: Hanning/Hamming y rectangular.

Análisis temporal localizado

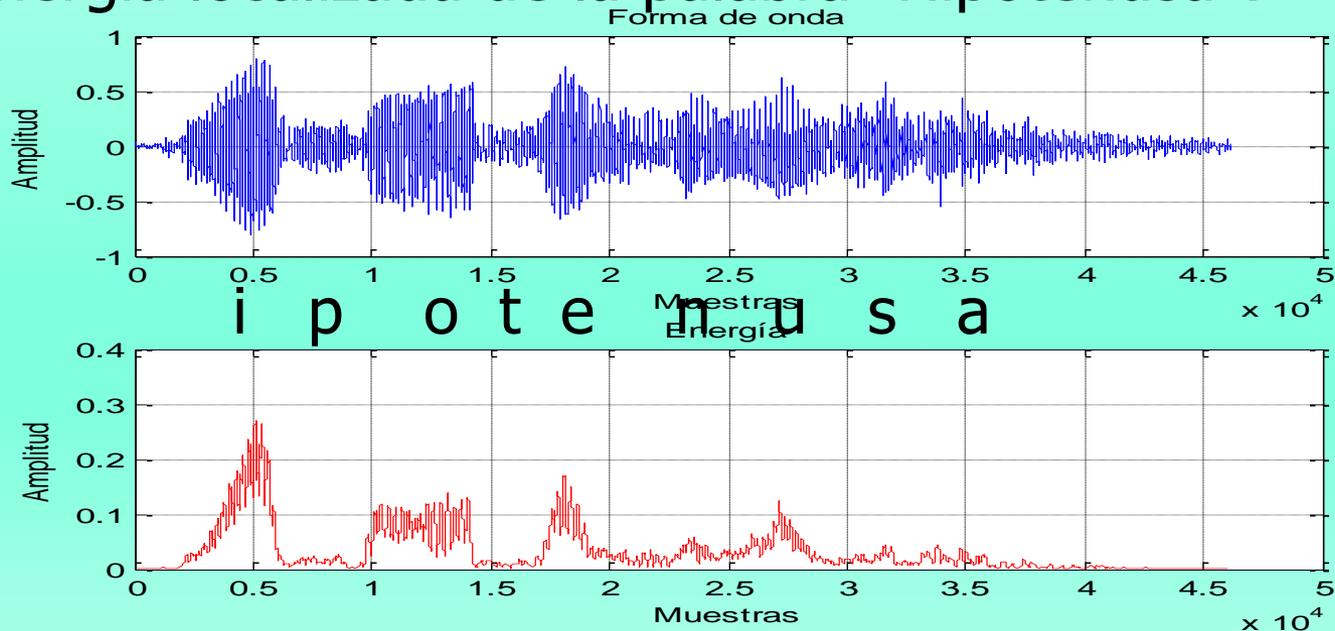
- Parámetros típicos que se suelen calcular:
 - Energía localizada (o en su defecto la magnitud)
 - Tasa de cruces por cero
 - Autocorrelación
 - Estimación de la frecuencia fundamental F_0 (*Pitch*)

Energía localizada

- $E[m]$: Energía localizada

$$E[m] = \sum_{n=-\infty}^{+\infty} [x[n]w[n-m]]^2 = \sum_{n=0}^{N-1} x[n]^2 w[n-m]^2$$

- Energía localizada de la palabra "Hipotenusa":



- $E[m]$: Energía localizada

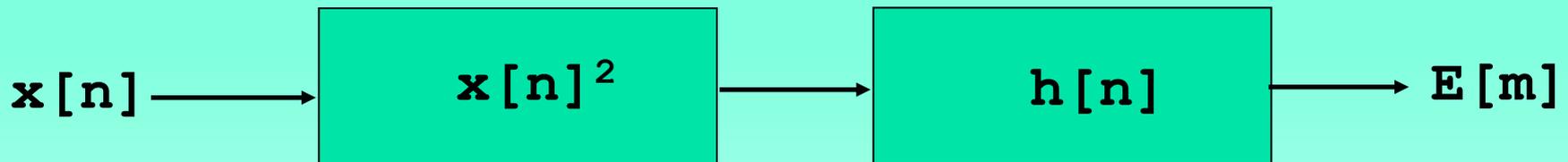
$$E[m] = \sum_{n=0}^{N-1} x[n]^2 w[n-m]^2$$

- Esta ecuación se puede interpretar como:

$$E[m] = \sum_{n=0}^{N-1} x[n]^2 h[n]$$

Siendo: $h[n] = w[n-m]^2$

- Esto a su vez se puede interpretar como:



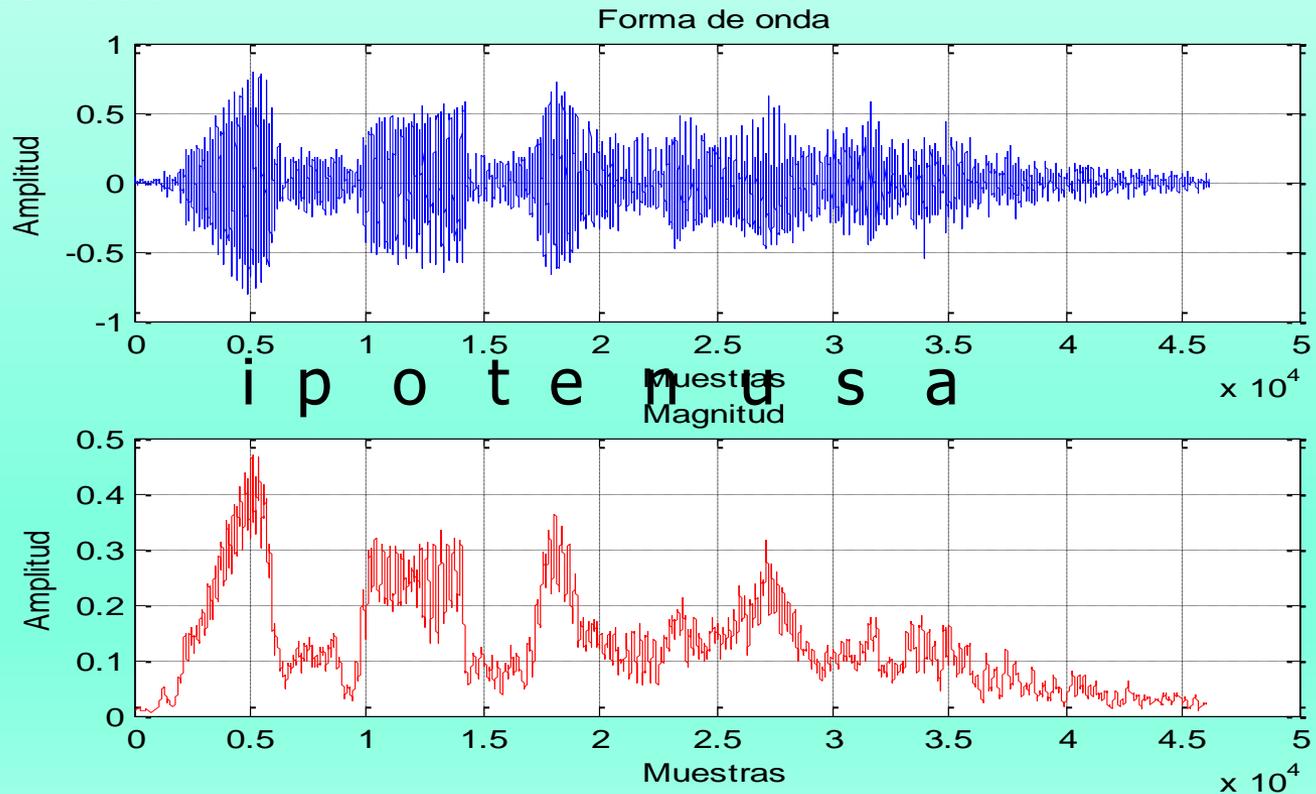
Magnitud

- $M[m]$: Magnitud

$$M[m] = \sum_{n=0}^{N-1} |x[n]| w[n-m]$$

- Es un parámetro alternativo a la energía
 - Menor complejidad
 - Menor margen dinámico
 - Muestras elevadas pueden desvirtuar el valor de la energía al ser elevadas al cuadrado

- Ejemplo del cálculo de la magnitud para la palabra “Hipotenusa”



Tasa de cruces por cero

- $T_{cc}[m]$: Tasa de cruces por cero

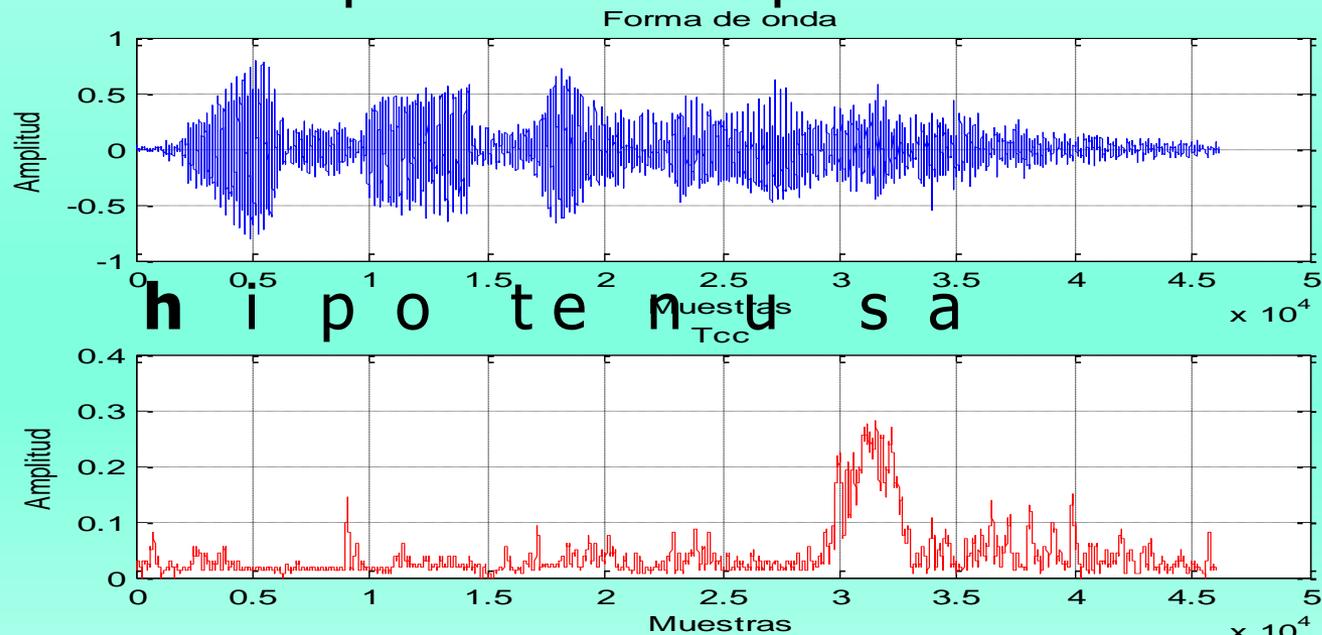
$$T_{cc}[m] = \frac{1}{N} \sum_n \frac{1}{2} |\text{sgn}(x[n]) - \text{sgn}(x[n-1])| w(m-n)$$

- Donde $\text{sgn}()$ es la función signo definida por:

$$\text{sgn}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

Tasa de cruces por cero

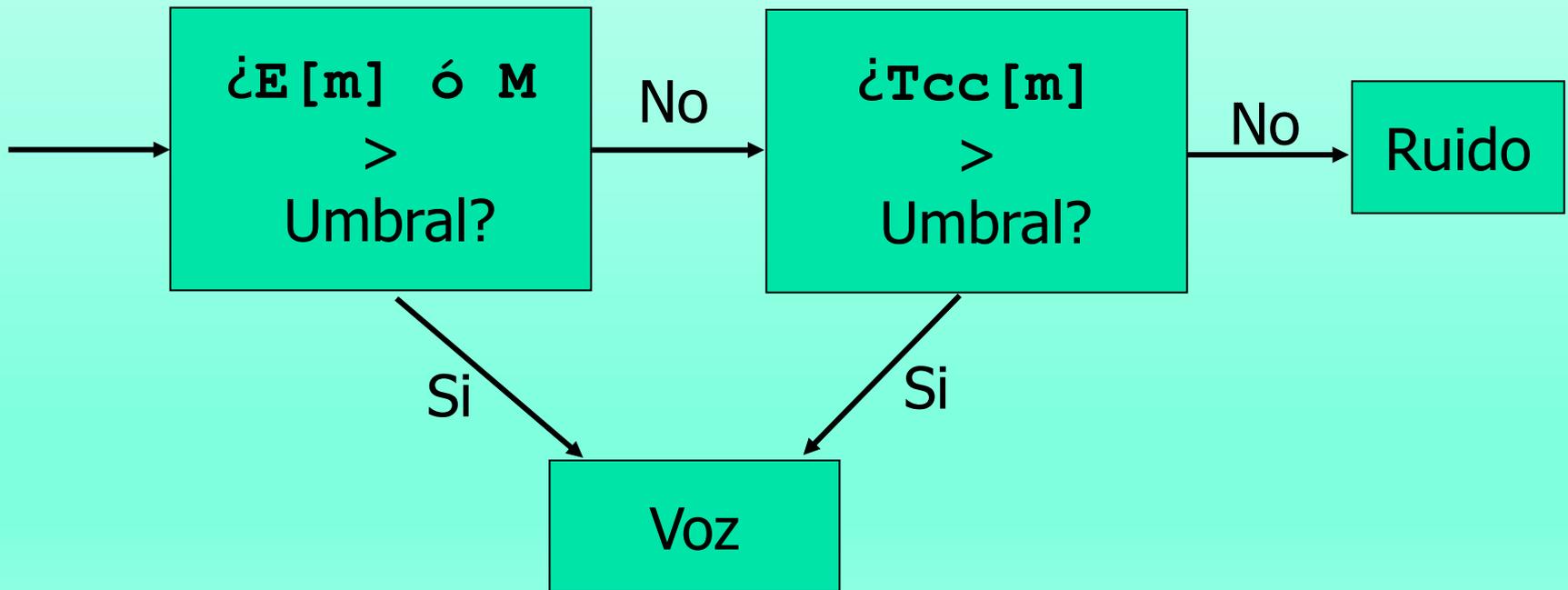
- Indica la relación entre la energía a baja y alta frecuencia.
- Las señales sonoras dan un tasa menor que las señales sordas.
- Tasa de cruces por cero de "Hipotenusa"



Aplicaciones E, M y Tcc

- Entre las principales aplicaciones se encuentran:
 - Clasificación de sonidos
 - Sonoros/Sordos, etc...
 - Detector de actividad (VAD: *Voice Activity Detector*)
 - Uso en codificación:
 - Ej. GSM: para reducir interferencias y ahorrar batería.
 - Uso en reconocimiento:
 - Mayor eficiencia y evitar reconocimientos erróneos.

- Detector de actividad:



Autocorrelación

- $R_m[k]$: Autocorrelación

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

$$k = 0, 1, 2, \dots, p.$$

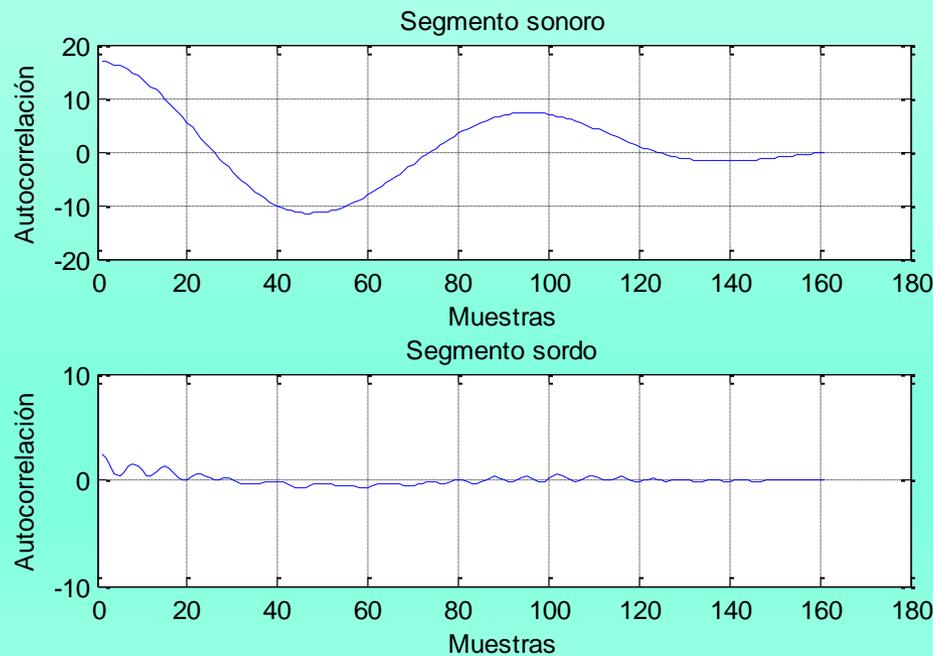
- Propiedades:

- Es una función par

- Tiene un máximo en $k=0$, i.e.: $|R_m[k]| \leq R_m[0]$

- $R_m[0] = \textit{Energía}$

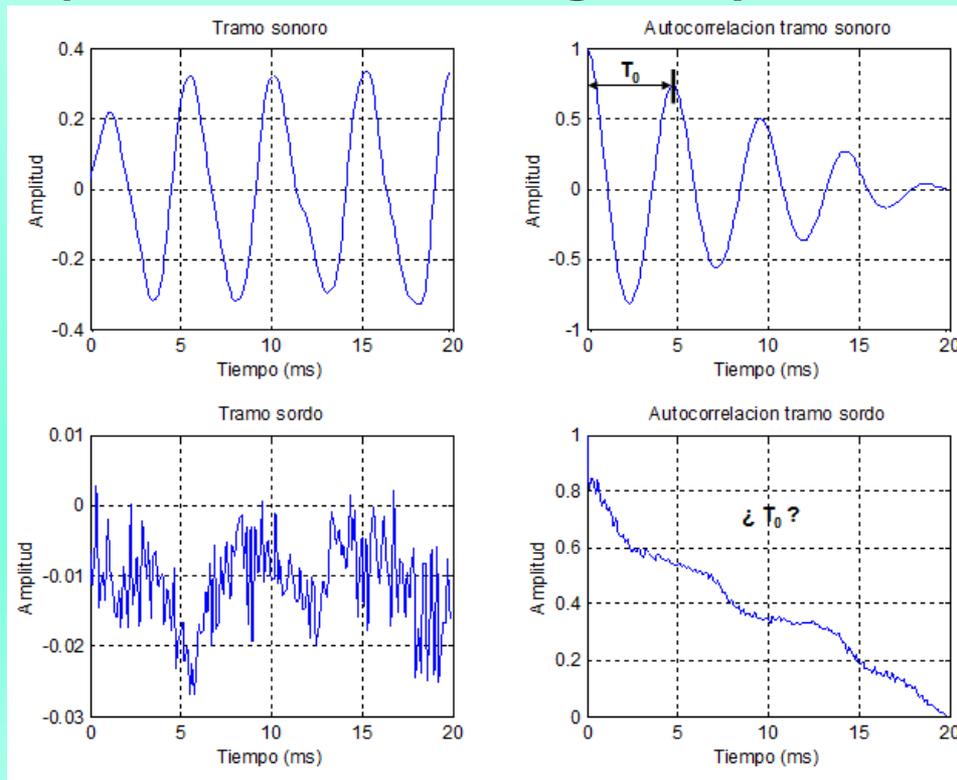
- Para desplazamientos de k igual al periodo de la señal la autocorrelación tiene máximos locales
- La autocorrelación de una señal periódica es periódica



- En una señal de voz:
 - Los máximos locales de la autocorrelación corresponden con el *pitch* (frecuencia fundamental, f_0) y los formantes del tracto vocal.

Estimación del Pitch

- A partir de la correlación
- Es el mayor máximo local de la autocorrelación (excluyendo el máximo global)

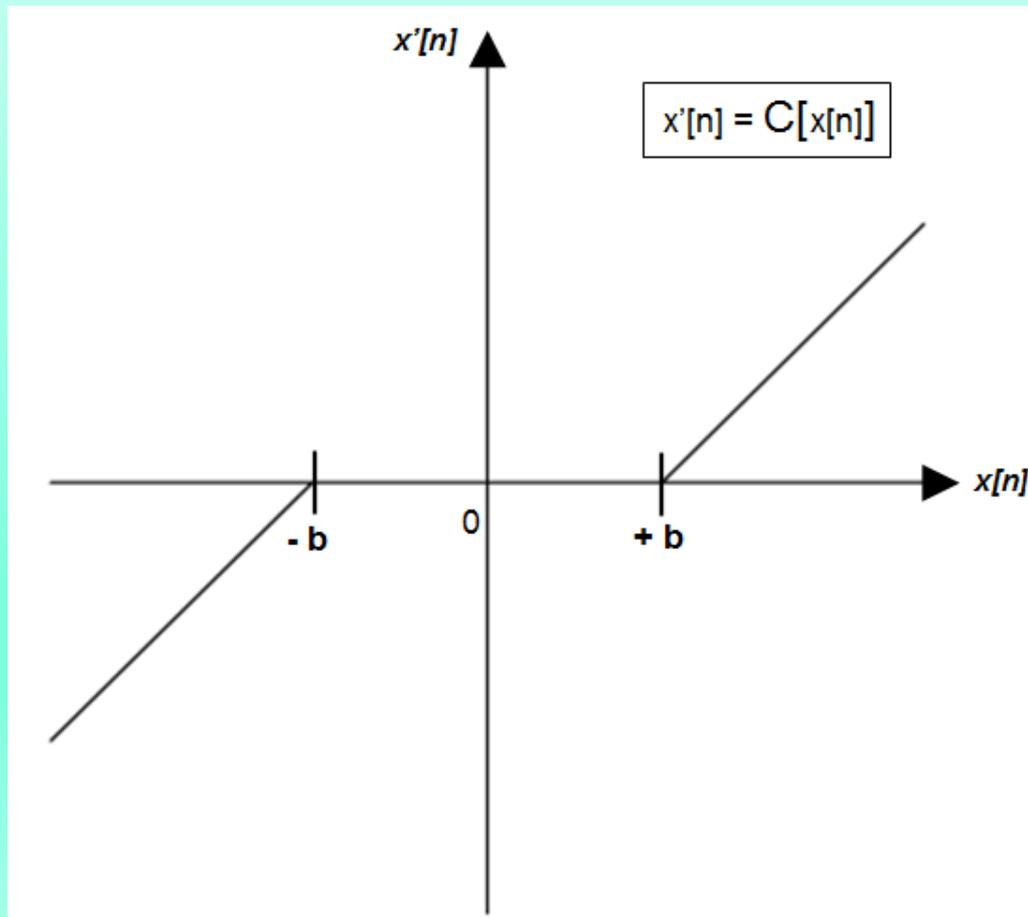


Segmento
Sonoro

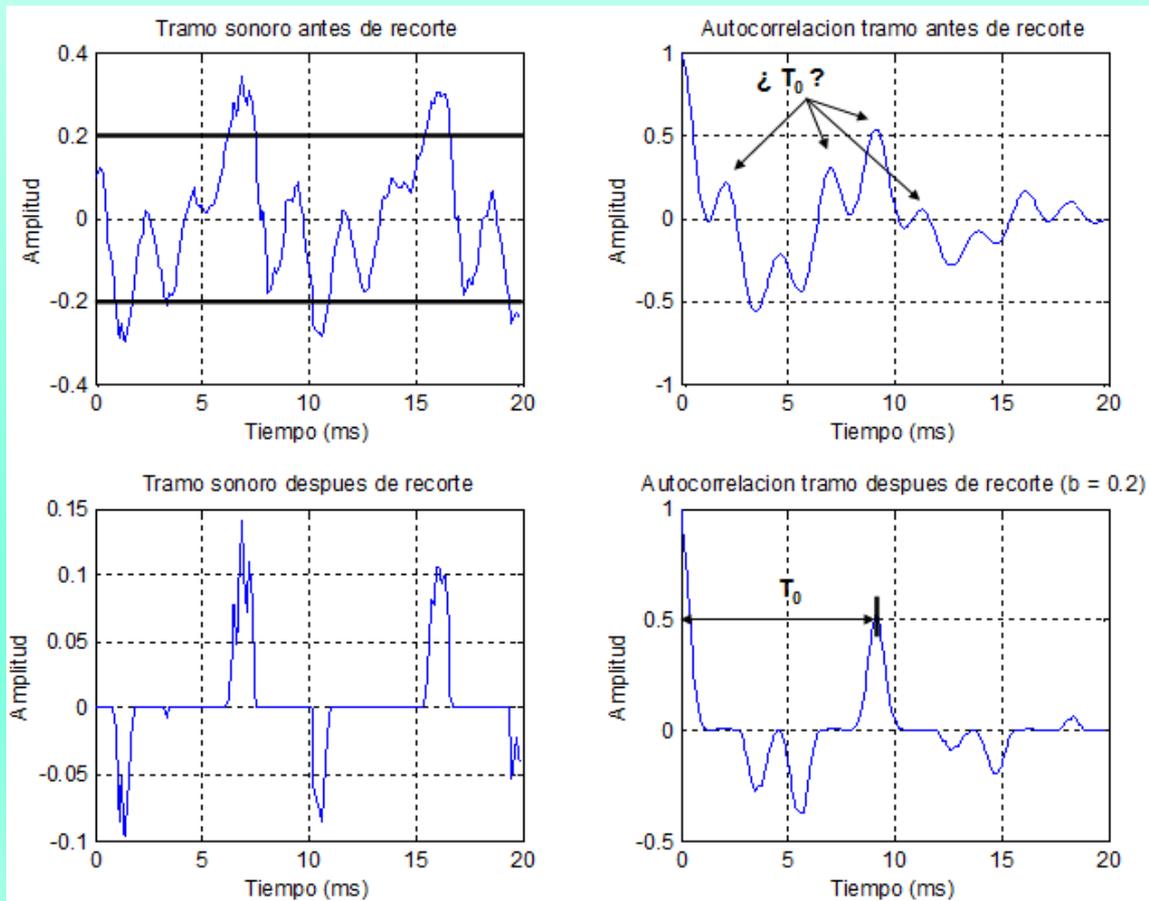
Segmento
Sordo

- Problema:
 - No siempre el mayor máximo local corresponde con el *pitch*
- Para facilitar su localización emplearemos una función de recorte
- Esta función eliminará toda la señal de entrada que no sobrepase un determinado umbral

- Función de recorte:



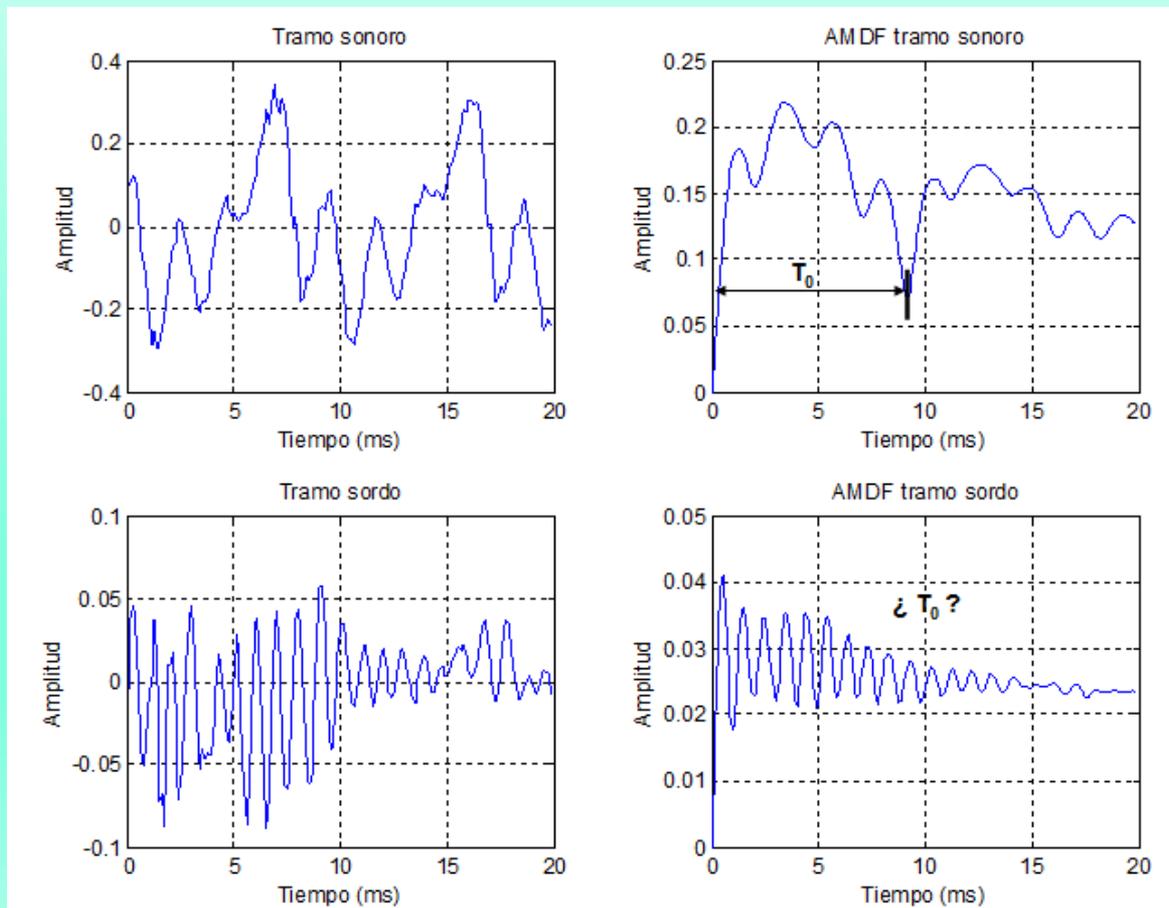
- Autocorrelación de la señal recortada



- *AMDF, Average Magnitude Difference Function*
 - Estima del pitch empleando la Magnitud en vez de la correlación
 - Menor complejidad y coste computacional
 - En este caso en vez de buscar máximos se deben buscar mínimos

$$AMDF[m, \tau] = \sum_{n=m+\tau}^{m+N-1} |s[n]w[n-m] - s[n-\tau]w[n-m-\tau]|$$

- *AMDF, Average Magnitude Difference Function*



Análisis localizado en frecuencia

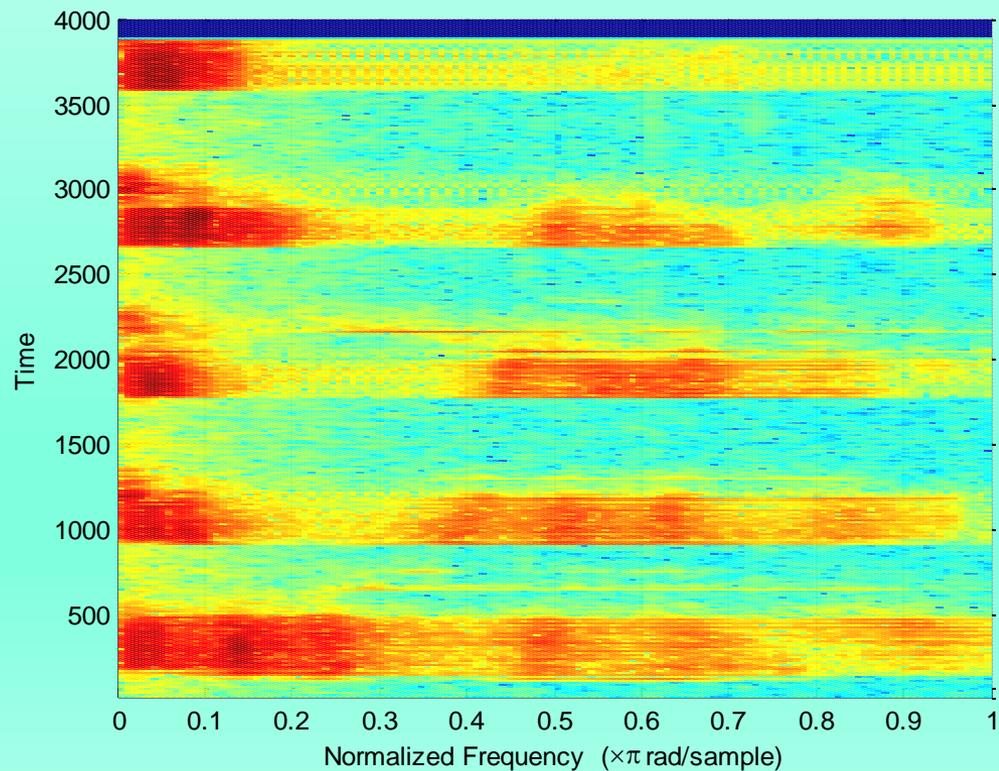
- Para realizar un análisis localizado en frecuencia basta con calcular la TF de un segmento de señal enventanado.

$$S(n, \omega) = \sum_{m=-\infty}^{\infty} s[m]w[n-m]e^{-j\omega m}$$

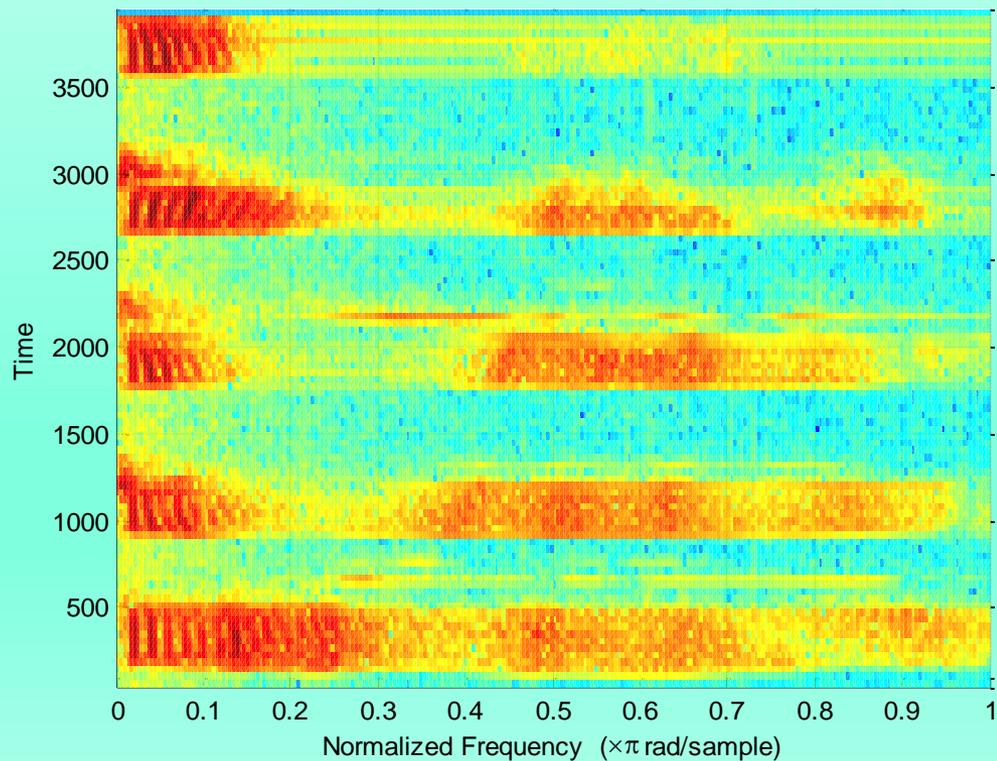
Espectrogramas

- También denominados Sonogramas
- Representan la evolución del espectro con el tiempo
- Estas variables son inversas
 - Al ganar resolución en una de ellas, la perdemos en la otra
- Tipos de espectrogramas:
 - Banda ancha
 - Banda estrecha

- Banda ancha (poca resolución en frecuencia)
 - Ventanas temporales cortas



- Banda estrecha (poca resolución en el tiempo)
 - Ventanas temporales largas



Análisis Homomórfico: Cepstrum

- Utilidad:
 - Permite separar la señal de excitación de la respuesta del filtro del tracto vocal
- Un segmento sonoro es la convolución entre:
 - La señal de excitación glotal $e[n]$
 - El filtro del tracto vocal $h[n]$

$$s[n] = e[n] * h[n]$$

- La convolución en el tiempo es una multiplicación en frecuencia

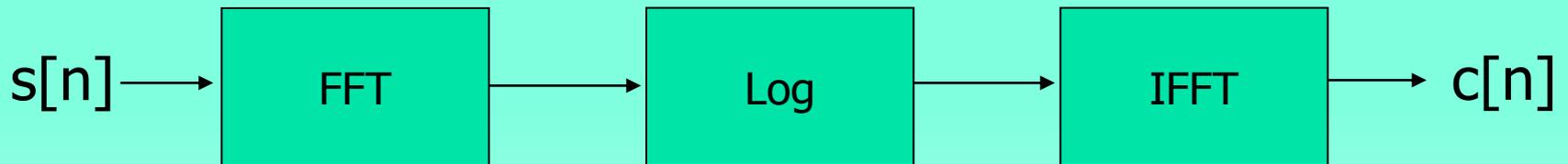
$$S[k] = E[k] \cdot H[k]$$

- Aprovechando las propiedades de los logaritmos:

$$\log(S[k]) = \log(E[k]) + \log(H[k])$$

- Si ahora regresamos al “tiempo”: Cepstrum

$$c[n] = IDFT[\log(S[k])]$$

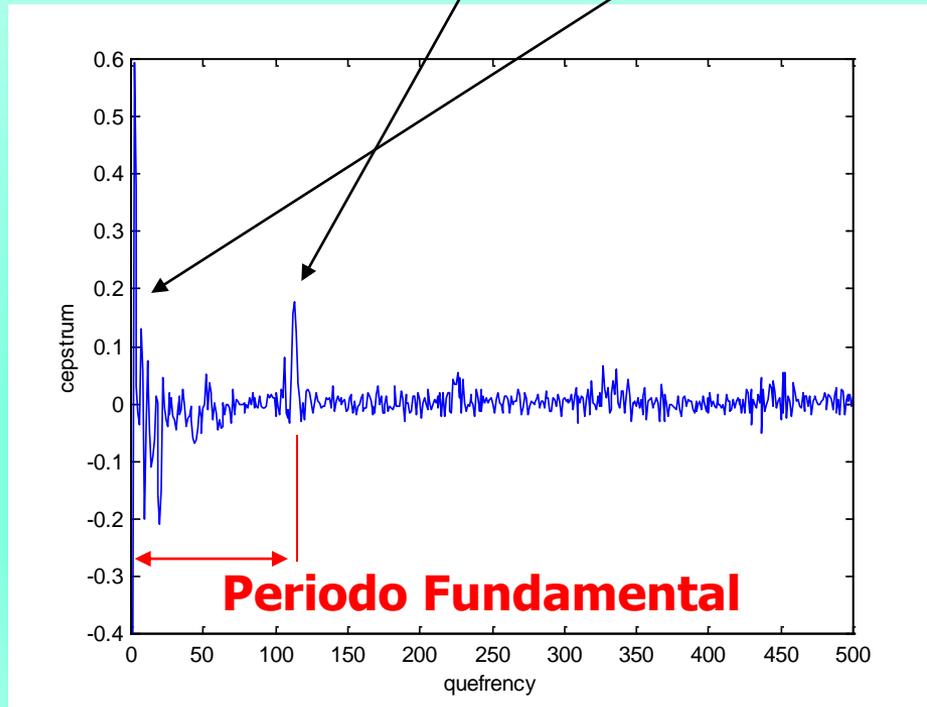


- El cepstrum puede ser real o complejo:
 - Cepstrum complejo: tomamos logaritmos del espectro completo (con la fase desenrollada, *unwrapped*)
 - Cepstrum real: sólo aplicamos el logaritmo al módulo del espectro
- El cepstrum complejo se puede deshacer, el real no al no contener información de fase
- Para voz se suele emplear el cepstrum real

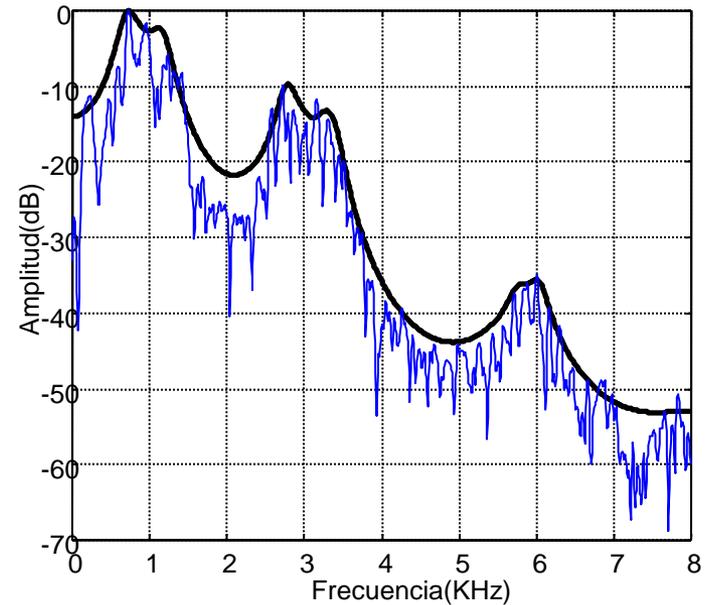
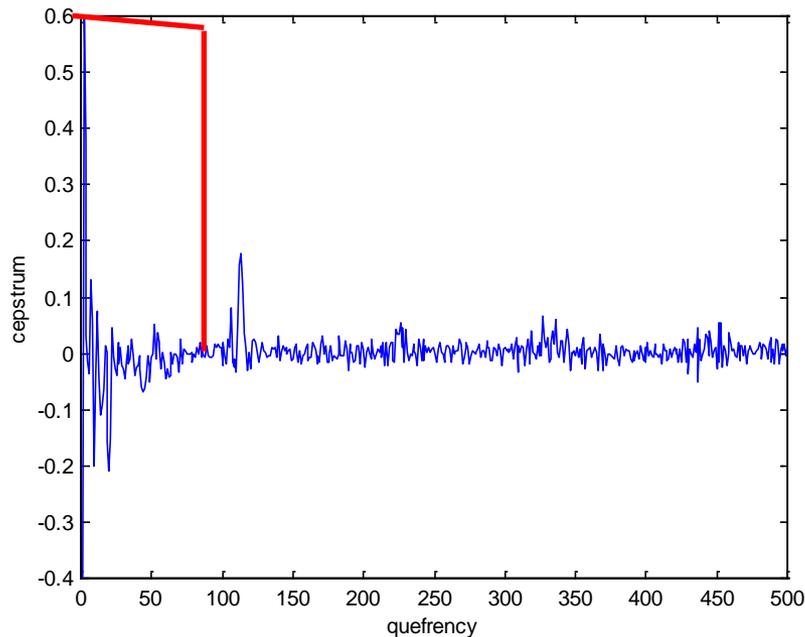
- La convolución se ha convertido en una suma:

$$s[n] = e[n] * h[n] \longrightarrow c[n] = c_e[n] + c_h[n]$$

c_e y c_h son separables



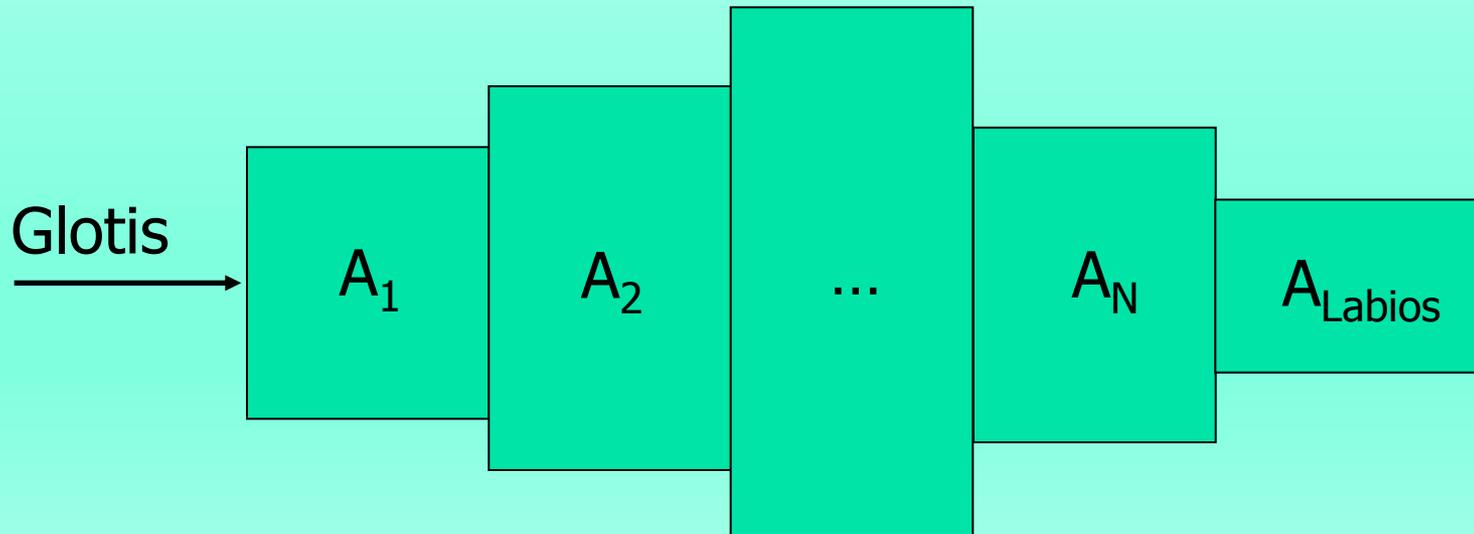
- Obtención de la envolvente espectral:
 - Una vez calculado el cepstrum
 - Extraemos c_h con una ventana
 - El espectro de c_h es la envolvente espectral



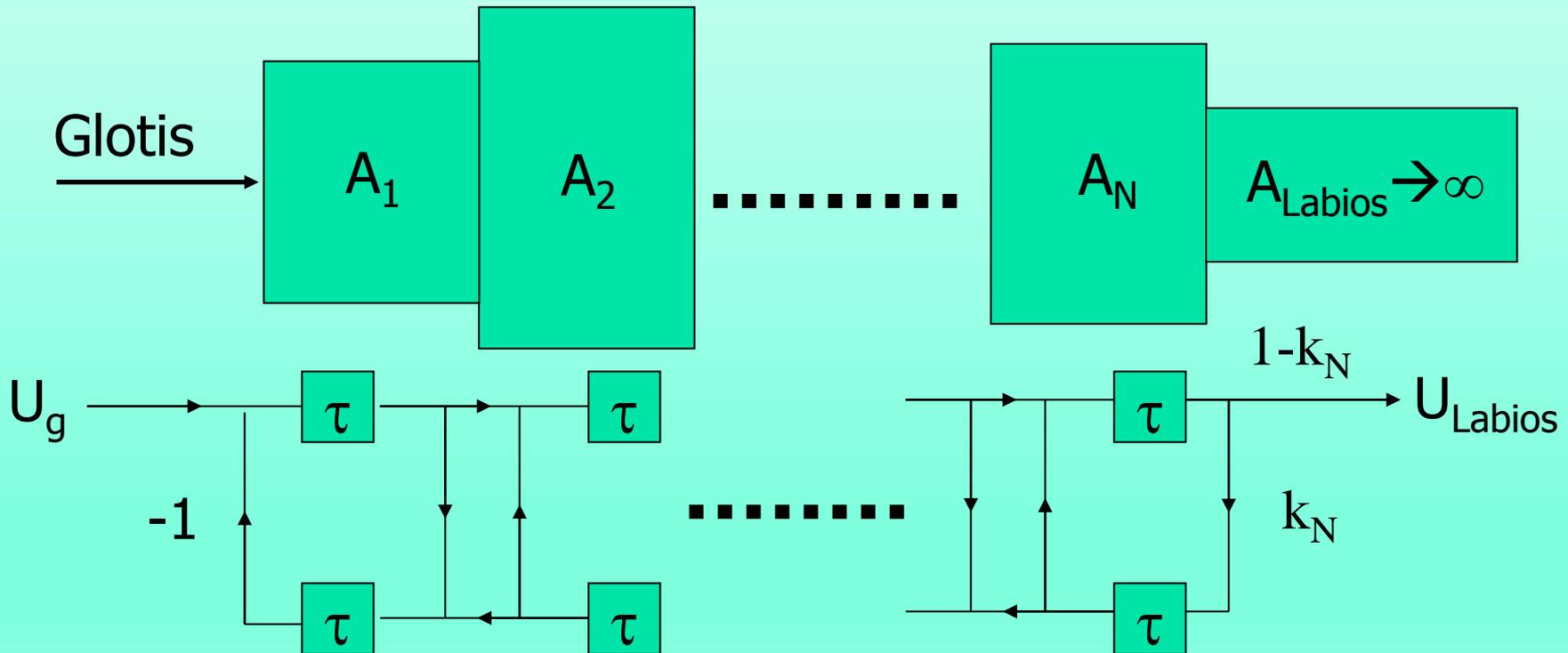
- Terminología empleada:
 - Spectrum → Cepstrum
 - Frecuency → Quefrequency
 - Filtering → Liftering
 - Analysis → Alanysis

Análisis de predicción lineal

- Modelo del tracto vocal:
 - Suponemos que el tracto vocal es una serie de tubos de sección variable sin pérdidas
 - Suponemos que el sonido se propaga como una onda plana a través de los tubos



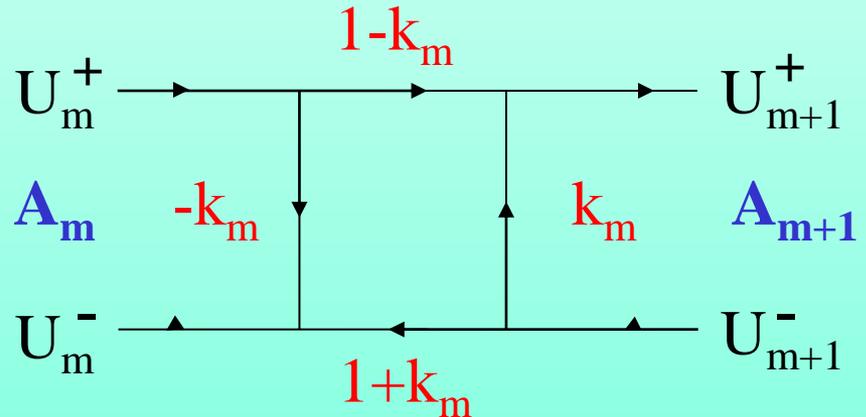
- Modelo del tracto vocal



- Estructura de filtro en celosía (*lattice*)
- τ tiempo de propagación para atravesar una sección

- Coeficientes de reflexión:

Interconexión
de secciones:



Cálculo de los coeficientes
de reflexión:

$$k_m = \frac{A_m - A_{m+1}}{A_m + A_{m+1}}$$

- Trabajando en tiempo discreto:
 - Si el periodo de muestreo $T = 2 \tau$ se puede demostrar que la respuesta en frecuencia del tracto vocal es un filtro todo polos
 - Los coeficientes a_k del filtro se pueden obtener a partir de los coeficientes de reflexión k_m (*Durbin*)

- Predicción lineal:
 - Vamos a intentar predecir el valor de $s[n]$ a partir de sus valores anteriores $s[n-1]$, $s[n-2]$, ..., $s[n-M]$
 - Es decir, $s[n]$ se puede calcular en función de sus muestras anteriores (podemos predecir su valor):
$$s[n] \leftrightarrow f \{s[n-1], s[n-2], \dots, s[n-M]\}$$
 - Si la función f es lineal: predicción lineal

- Cálculo de la predicción de $s[n]$:

$$\hat{s}[n] = a_1s[n-1] + a_2s[n-2] + \dots + a_Ps[n-P]$$

- Coeficientes de predicción:

$$\{a_1, a_2, \dots, a_P\}$$

- Error de predicción:

$$e[n] = s[n] - \hat{s}[n] = s[n] - \sum_{i=1}^P a_i s[n-i]$$

- Cálculo de los coeficientes de predicción:
 - Son aquellos que minimizan el error de predicción (la energía del error de predicción)

$$E_p = \sum e[n]^2 = \sum \left\{ s[n] - \sum_{i=1}^P a_i s[n-i] \right\}^2$$

- Minimizar: Para cada a_k derivar e igualar a 0

$$\frac{\partial E_p}{\partial a_k} = 0 \quad k = 1, \dots, P$$

- Obtenemos un sistema de P ecuaciones con P incógnitas

- Cálculo de los coeficientes de predicción:

$$E_p = \sum \left\{ s[n] - \sum_{i=1}^P a_i s[n-i] \right\}^2$$

$$\frac{\partial E_p}{\partial a_k} = 0 \quad \longrightarrow \quad 2 \sum \left\{ s[n] - \sum_{i=1}^P a_i s[n-i] \right\} (-s[n-k]) = 0$$

$$\sum \{ s[n] s[n-k] \} = \sum_{i=1}^P a_i \left\{ \sum s[n-i] s[n-k] \right\}$$

$$R_s[|k|]$$

$$R_s[|k-i|]$$

- Cálculo de los coeficientes de predicción:

$$R_s[|k|] = \sum_{i=1}^P a_i R_s[|k-i|] \quad \text{para } k = 1..P$$

- Sistema de ecuaciones:

$$R_s[1] = a_1 R_s[0] + a_2 R_s[1] + \dots + a_p R_s[P-1]$$

$$R_s[2] = a_1 R_s[1] + a_2 R_s[0] + \dots + a_p R_s[P-2]$$

...

$$R_s[P] = a_1 R_s[P-1] + a_2 R_s[P-2] + \dots + a_p R_s[0]$$

- En forma matricial:

Ecuaciones de *Yule-Walker*

$$\begin{bmatrix} R_s[1] \\ R_s[2] \\ \dots \\ R_s[P] \end{bmatrix} = \begin{bmatrix} R_s[0] & R_s[1] & \dots & R_s[P-1] \\ R_s[1] & R_s[0] & \dots & R_s[P-2] \\ \dots & \dots & \dots & \dots \\ R_s[P-1] & R_s[P-2] & \dots & R_s[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix}$$

$$r = R \cdot a \quad \longrightarrow \quad a = R^{-1} \cdot r$$

R es una matriz *Toeplitz*

- Algoritmo de *Durbin*:
 - Solución recursiva para calcular los coeficientes a_k aprovechando que R es *toeplitz*.
- Inicio: $E^{(0)} = r[0]$

- Recursión: $i=1, \dots, P$

Coef. Reflexión (PARCOR) $\longrightarrow k_i = \frac{1}{E^{(i-1)}} \left(r[i] - \sum_{j=1}^{i-1} a_j^{(i-1)} r[i-j] \right)$

Coef. LPC $\longrightarrow \begin{cases} a_i^{(i)} = k_i \\ a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \end{cases} \quad j = 1, \dots, i-1$

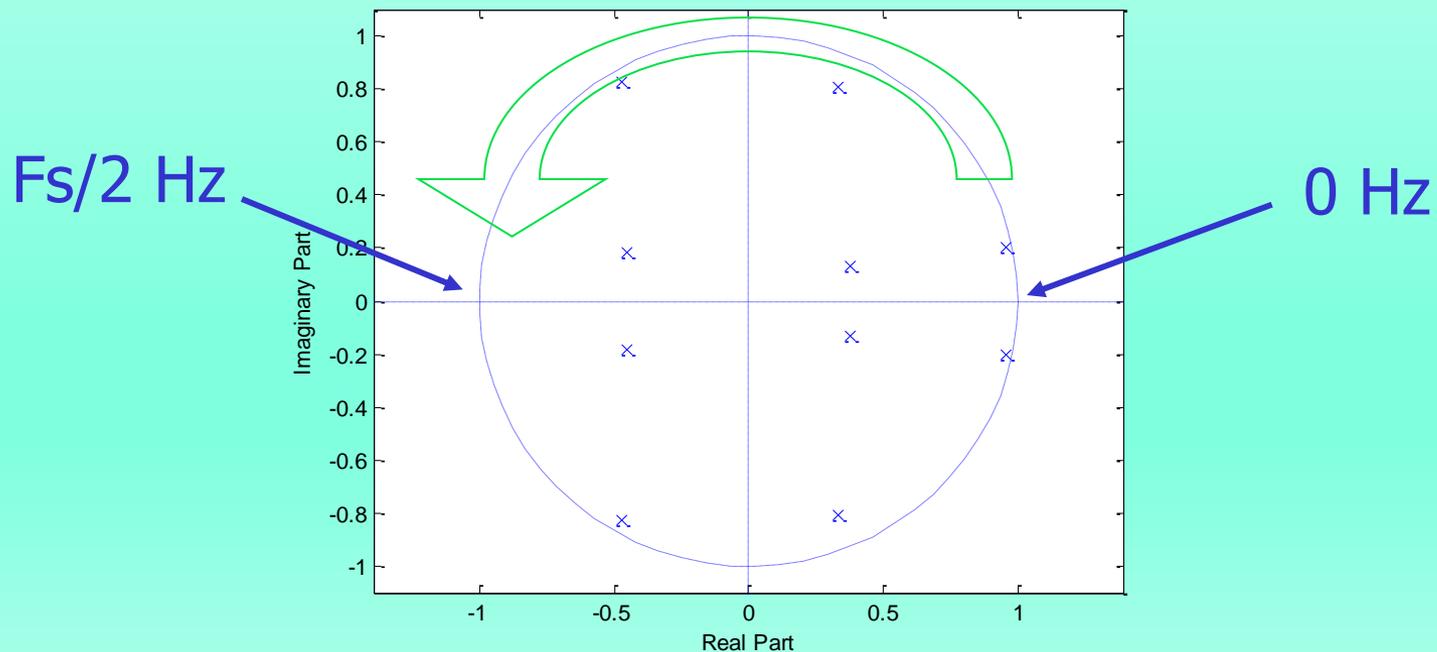
Energía residual $\longrightarrow E^{(i)} = (1 - k_i^2) E^{(i-1)}$

- Algoritmo de *Durbin*:
 - Calcula los coeficiente de reflexión (PARCOR)
 - Calcula los coeficientes de predicción lineal a partir de los de reflexión
 - El filtro resultante siempre es estable:
 - $|k_m| < 1$
- Filtro obtenido: IIR todo polos

$$H(z) = \frac{b_0}{1 + \sum_{k=1}^P a_k z^{-k}}$$

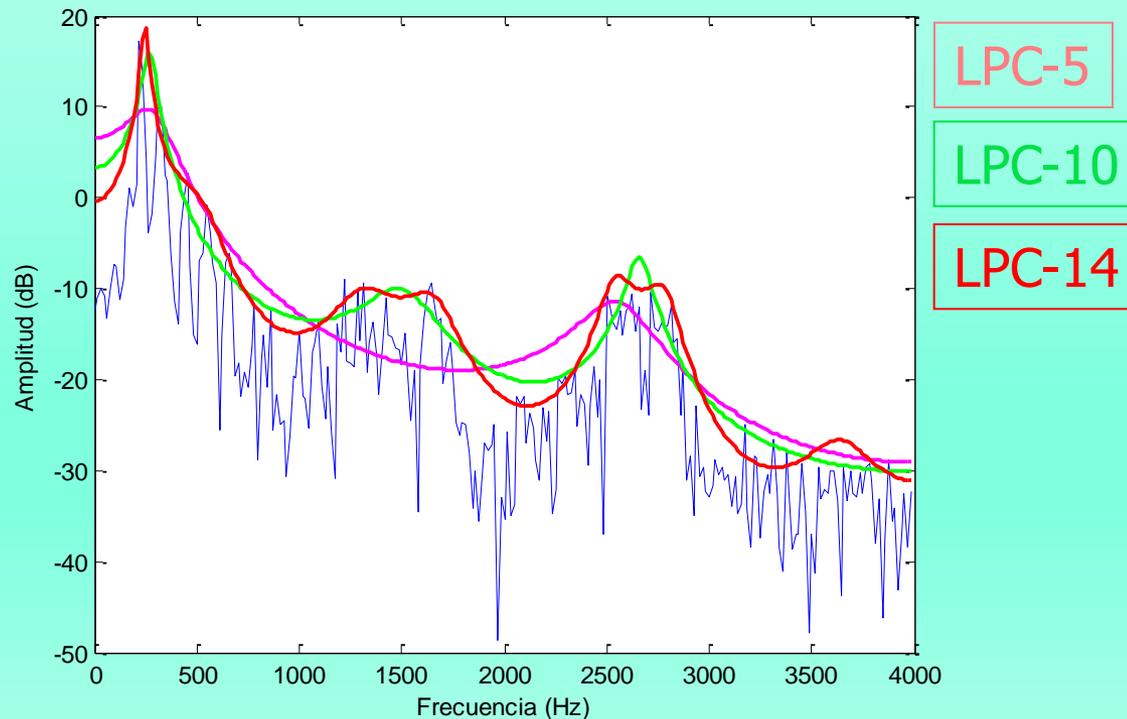
$$H(z) = \frac{b_0}{1 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_p z^{-p}}$$

- Cálculo de las frecuencias de los formantes:
 - A partir de los a_k calcular las raíces del polinomio
 - El cálculo de estas raíces debe hacerse de forma aproximada por métodos numéricos ya que no puede hacerse de forma analítica para polinomios grandes

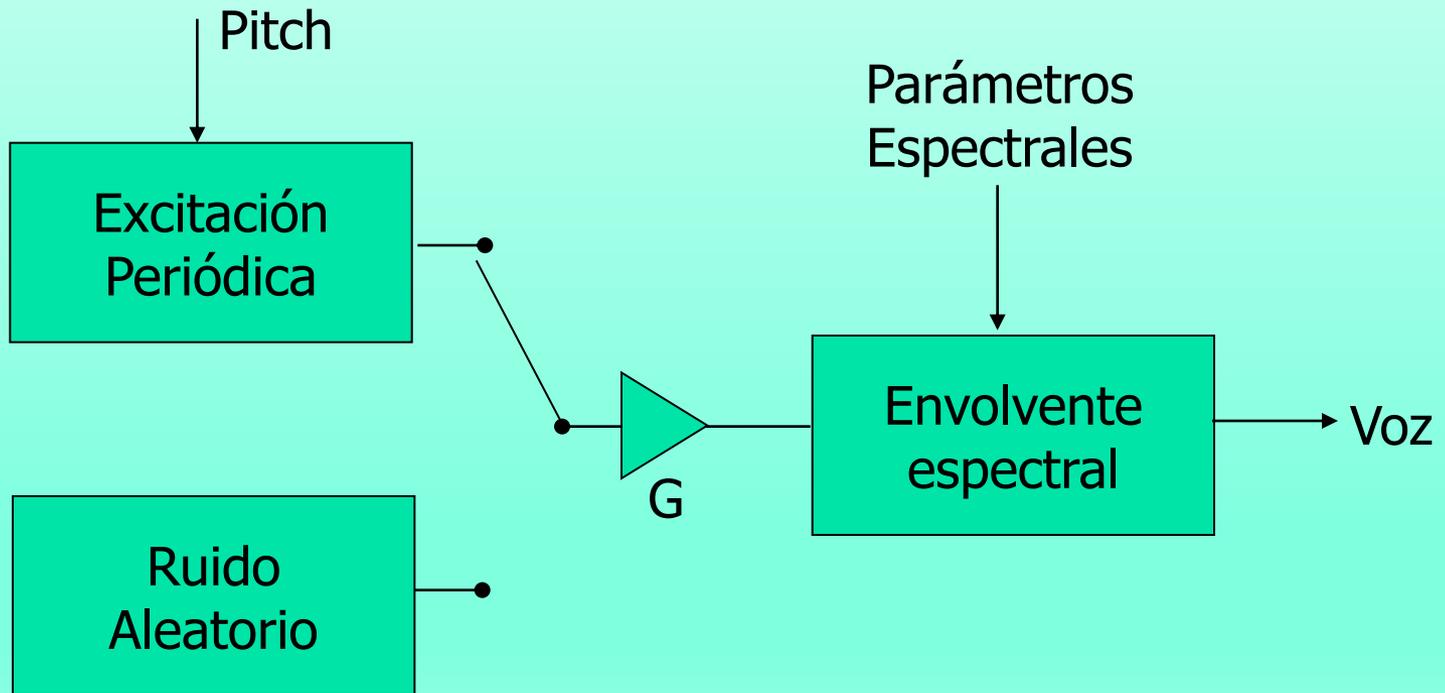


- Orden de predicción:

- Un par de polos complejos conjugados por cada formante
- Añadir dos o tres polos más
- En general P suele estar entre 10 y 14 coeficientes

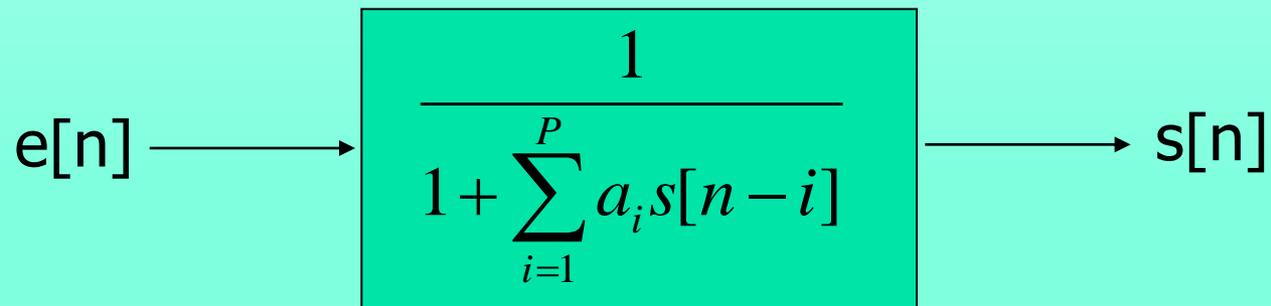


- Modelo de producción de voz:

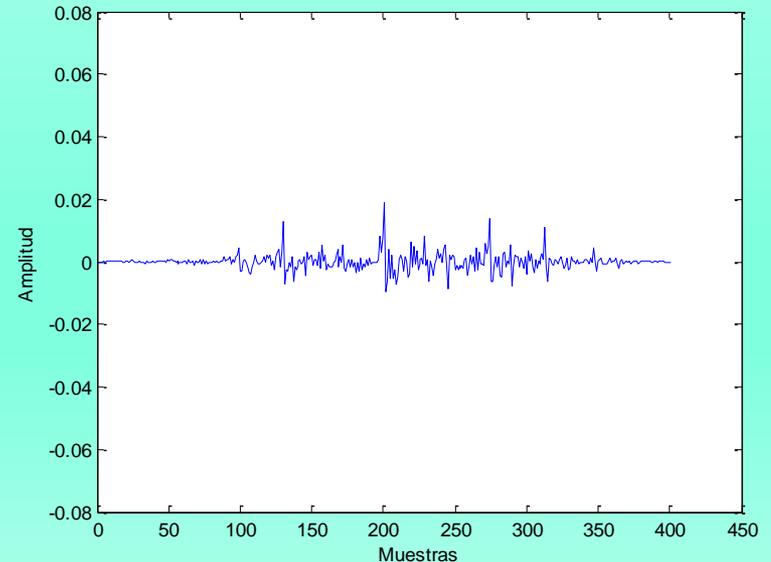
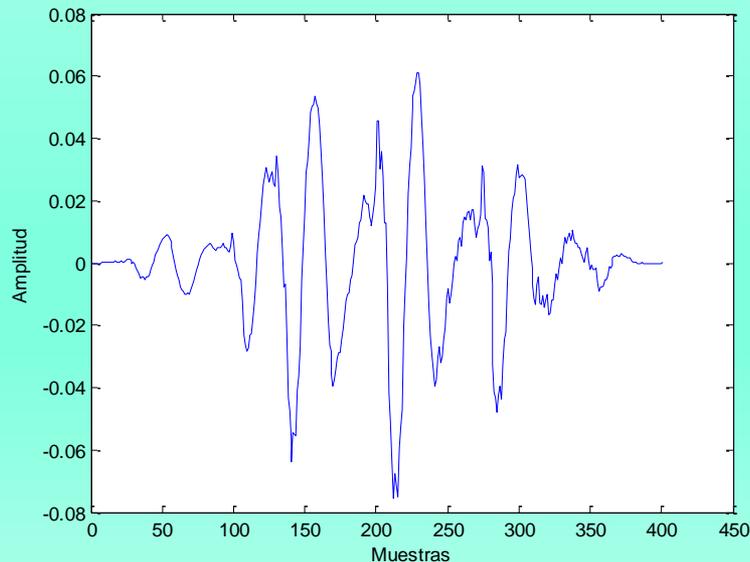
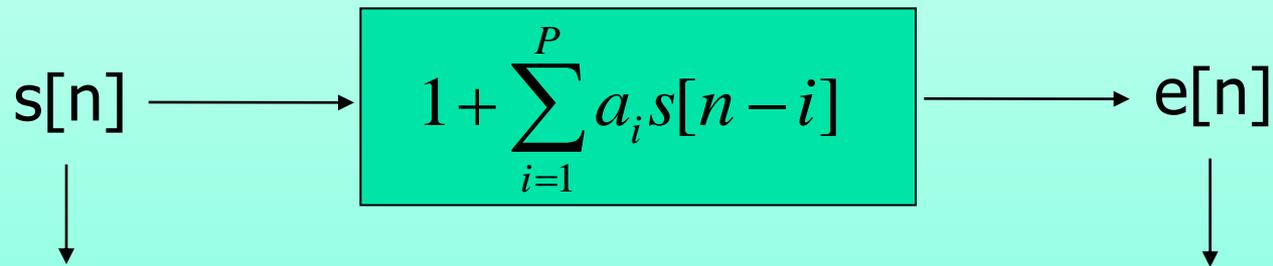


- A partir del error de predicción y del filtro LPC podemos obtener $s[n]$:

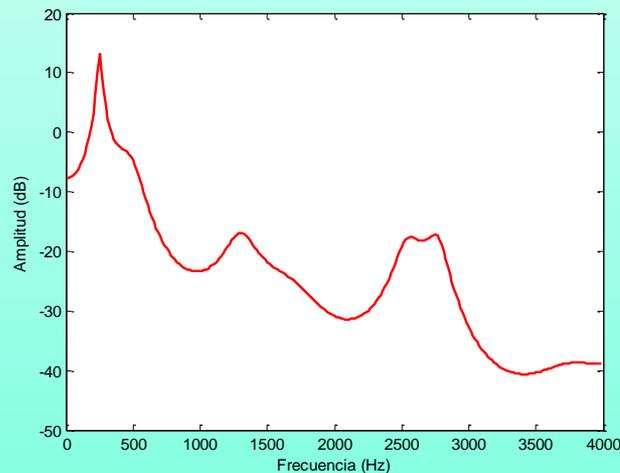
$$e[n] = s[n] - \sum_{i=1}^P a_i s[n-i] \quad \longrightarrow \quad s[n] = e[n] + \sum_{i=1}^P a_i s[n-i]$$



- Con el filtro LPC inverso y la señal de voz podemos obtener la señal de error:

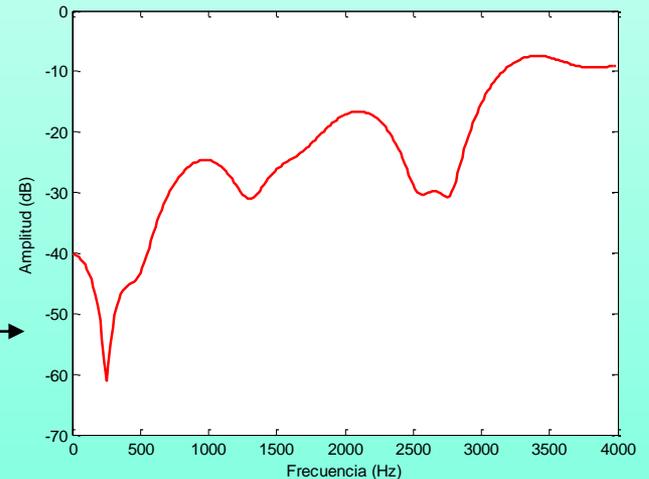


- Filtros LPC y LPC inverso:



← $H(z)$

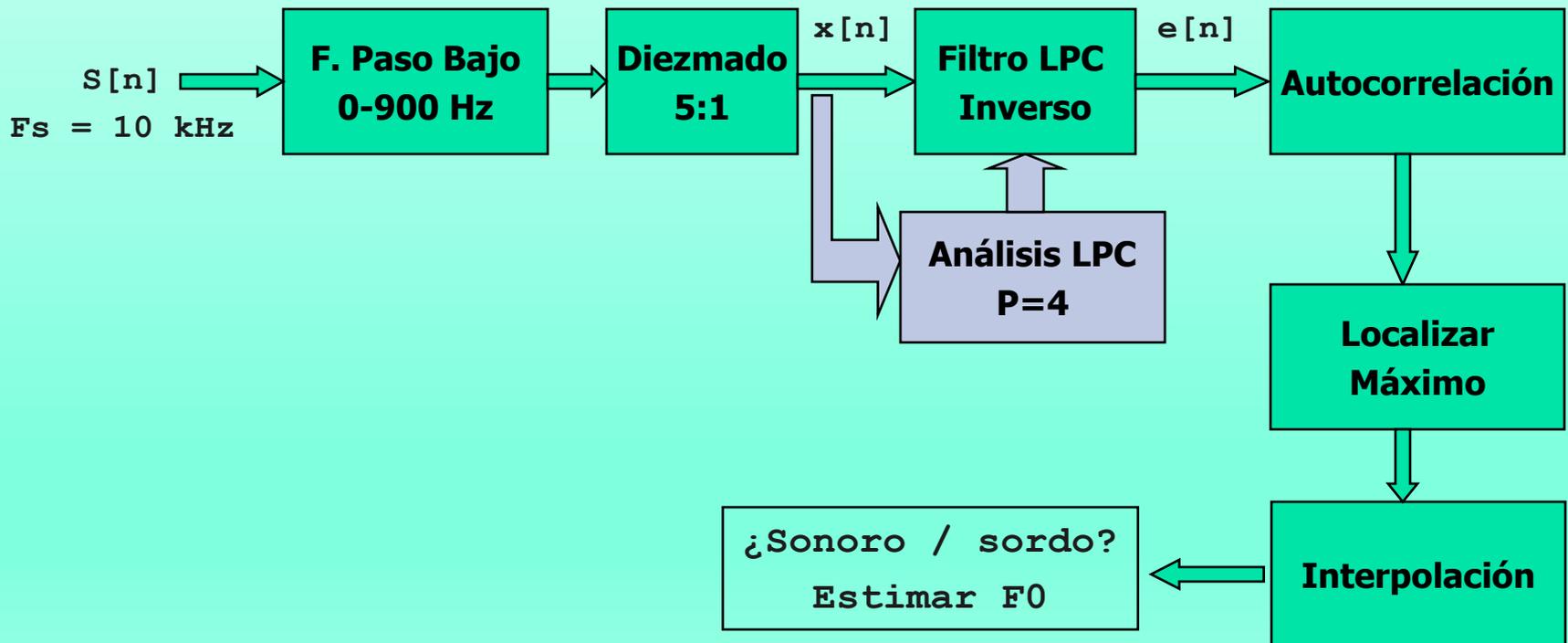
$\frac{1}{H(z)}$ →



- Al pasar $s[n]$ por el filtro LPC inverso obtenemos $e[n]$
- $e[n]$ además de ser la señal de error es la señal de excitación del modelo de producción de voz

Método SIFT, estimación del Pitch

- *Simplified Inverse Filtering Technique*, Markel 1972.



- Filtrar paso bajo con $f_c = 900\text{Hz}$.
- Esto nos permite reducir F_s de 10 kHz a 2 kHz.
 - Desechamos 4 de cada 5 muestras.
- Realizamos un análisis LPC de orden 4.
 - No es necesario más: hasta 1000Hz como máximo 2 formantes.
- Procesamos $x[n]$ con el filtro inverso LPC.
 - Obtenemos $e[n]$ que será la señal de excitación.
- Calculamos la autocorrelación de $e[n]$.
 - Localizamos el mayor valor dentro del rango de *pitch* probables.
- Para obtener mayor resolución en la estima del *pitch*, interpolamos la autocorrelación en la región del máximo.
- Si el máximo obtenido (normalizado por $R[0]$) no supera un umbral, suponer que el segmento es sordo.

Análisis espectral localizado

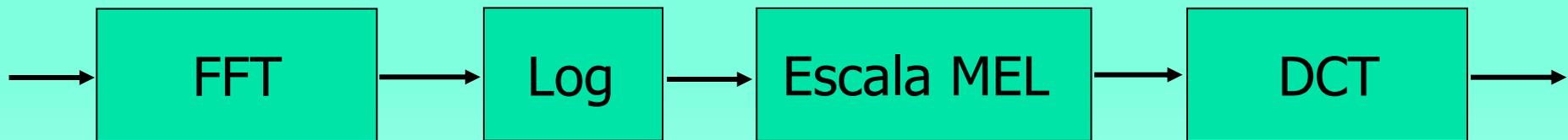
3.5.1.- Conceptos de percepción auditiva

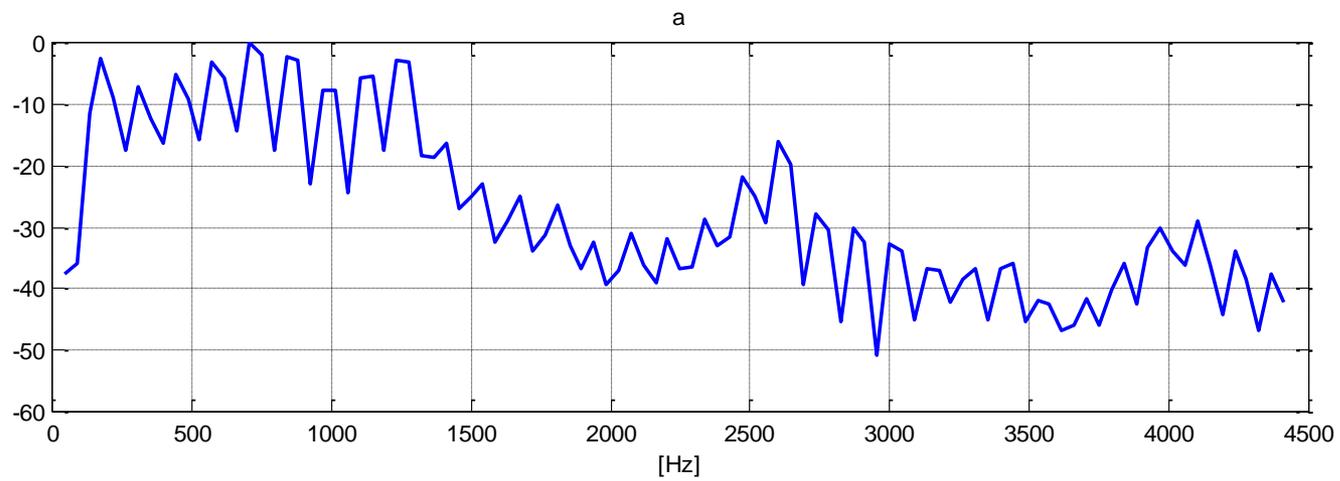
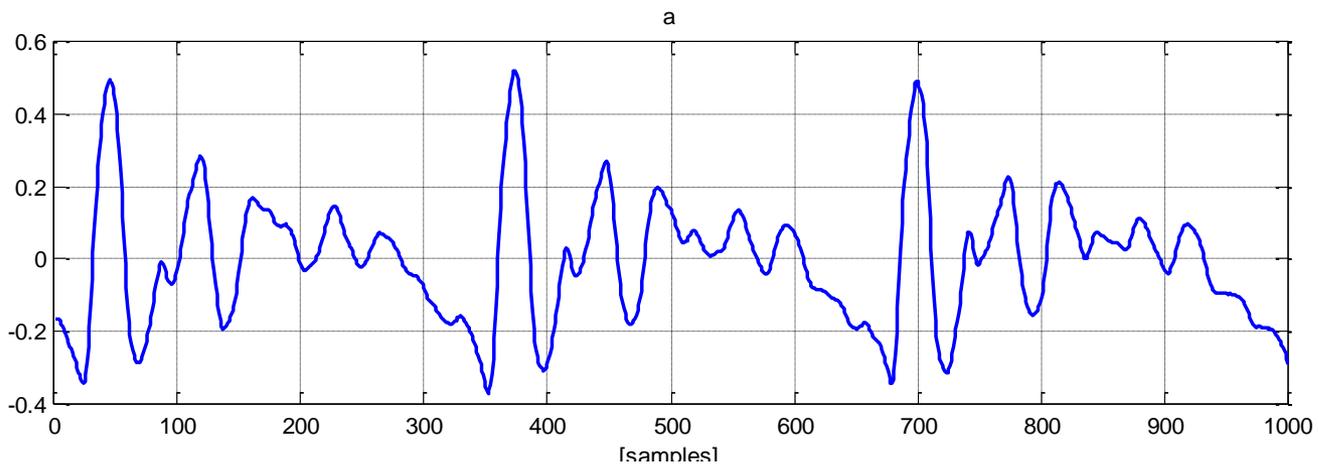
- MEL: Escala de frecuencias de distribución no lineal que responde al mecanismo de percepción auditiva
- Con esta escala medimos la frecuencia en MELs, es la frecuencia percibida aparente.
- Conversión de Hz a MELs

$$m = 1125 \cdot \log(0.0016f + 1)$$

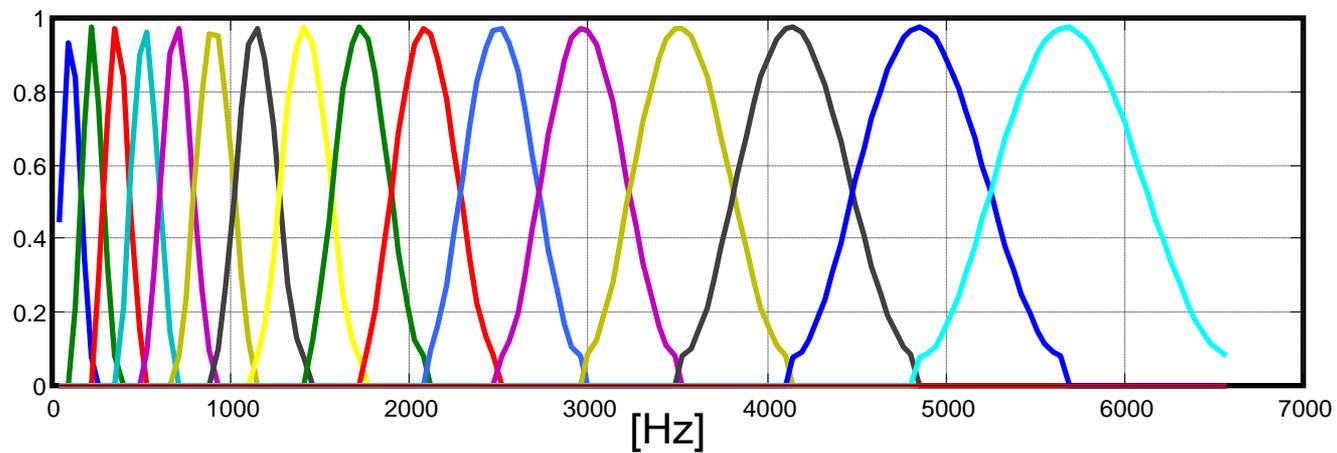
MEL-Frequency Cepstrum (MFCC)

- Coeficientes cepstrales derivados del análisis sobre la escala MEL
 - Calculamos el espectro
 - Calculamos el Log del módulo (cepstrum real)
 - Aplicamos la escala MEL
 - Agrupamos frecuencias en bandas críticas
 - Calculamos la DCT

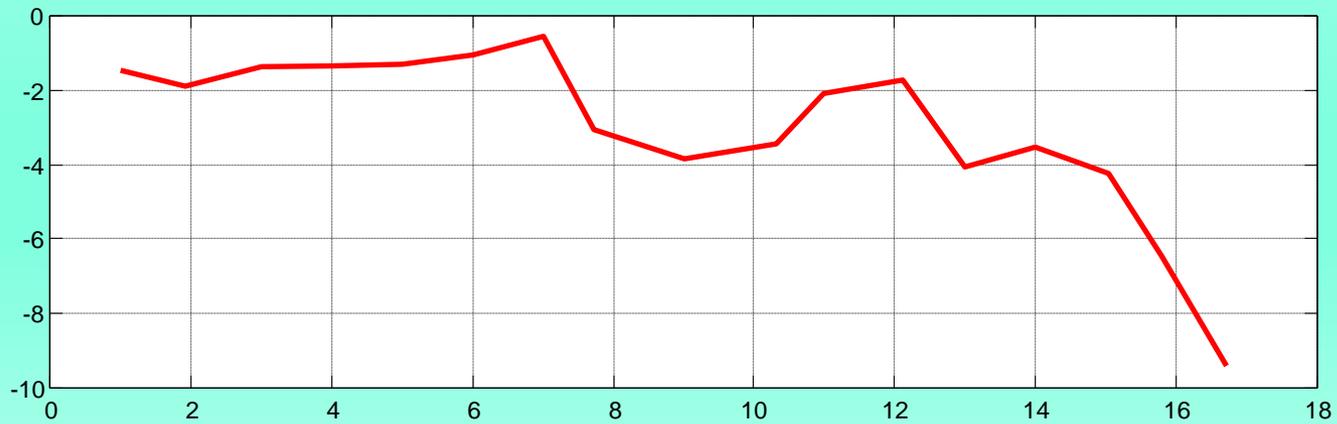




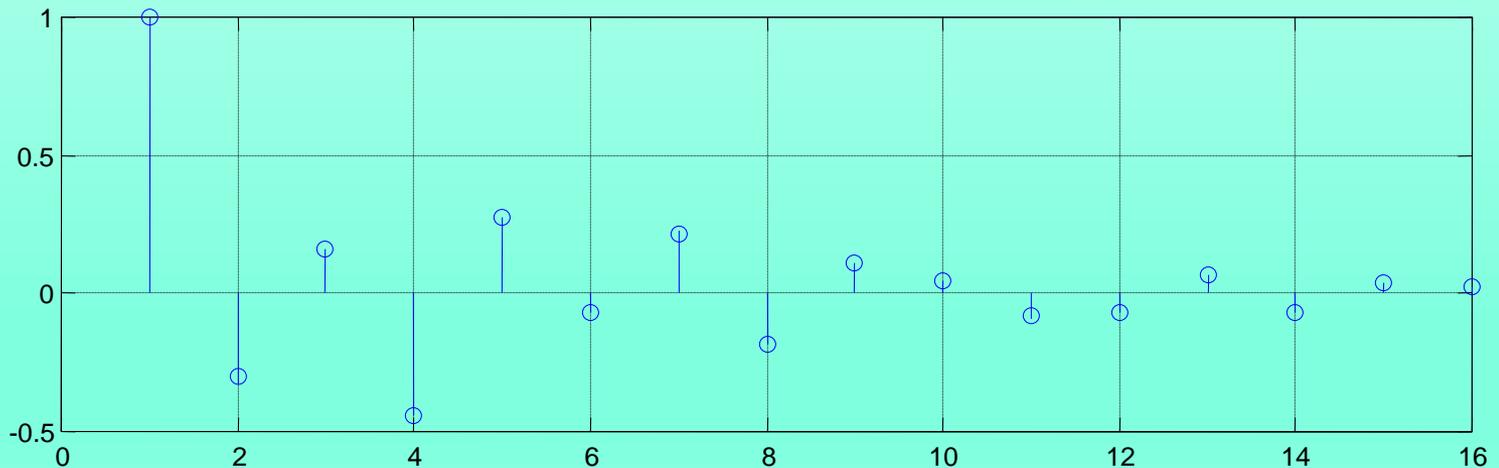
Banco de filtros



Espectro suavizado



- Cepstrum obtenido:
 - El número de coeficientes resultante es muy inferior
 - El cepstrum obtenido es una aproximación



Cepstrum LPC (LPCC)

- Es posible obtener los coeficientes cepstrales a partir de los coeficientes LPC
- Obtendremos el cepstrum de una señal suavizada
- No es necesario calcular el espectro

$$c(1) = -a_1$$

$$c(n) = -a_n - \sum_{m=1}^{n-1} \left(1 - \frac{m}{n}\right) a_m c(n-m) \quad n = 2..P$$

$$c(n) = -\sum_{m=1}^P \left(1 - \frac{m}{n}\right) a_m c(n-m) \quad n > P$$

Otros parámetros

- Existen multitud de representaciones distintas de los parámetros vistos
- Unos parámetros se pueden obtener a partir de los otros
- El empleo de unos u otros parámetros es indistinto en cuanto a mejoras en la síntesis/reconocimiento
- La elección entre unos u otros se debe principalmente a:
 - Robustez que ofrecen frente a fallos
 - Tasa binaria mínima requerida

- Coeficientes PARCOR:
 - *PARTial autoCORrelation coefficients.*
 - Se calculan como paso intermedio en el algoritmo de *durbin*.
 - Son los coeficientes de Reflexión ya vistos.

- Relación de áreas / Coefs. PARCOR

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

- LAR: *Log Area Ratios*

$$L_i = \log \left(\frac{1 - k_i}{1 + k_i} \right)$$

- Coeficientes LSF / LSP:
 - *Line Spectral Frequencies / Line Spectral Pairs*
 - Permiten una representación distinta de los coeficientes LPC
 - El filtro inverso LPC, $A(z)$, se puede descomponer en:

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1})$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1})$$

- Donde $P(z)$ representa la respuesta del tracto vocal con la glotis cerrada, y $Q(z)$ con la glotis abierta.

- $A(z)$ tiene raíces dentro de la circunferencia unidad
- $P(z)$ y $Q(z)$ sólo tienen raíces sobre la circunferencia
- $P(z)$ es un polinomio simétrico y $Q(z)$ antisimétrico
- Las raíces de $P(z)$ y $Q(z)$ se encuentran de forma alternada en frecuencia
- Cálculo de las raíces:
 - Tomar $z = \exp(j\omega)$ y evaluar $P(z)$ y $Q(z)$ en una malla de puntos entre 0 y π .
- Recuperación de $A(z)$:

$$A(z) = \frac{1}{2} [P(z) + Q(z)]$$

- Problemas de usar los coeficientes LPC:
 - El error de cuantificación es problemático, el filtro se puede hacer inestable
 - Se comportan muy mal al intentar interpolarlos
- Ventajas de usar LSF/LSP:
 - Son más robustos en cuanto a errores de cuantificación
 - El filtro permanece estable
 - Al ser una representación en frecuencia, un error solo altera un pequeño rango de frecuencias

Proceso de obtención de parámetros

- Pasos a realizar:
 - Pre-énfasis de la trama
 - Enventanado con solapamiento
 - Cálculo de la autocorrelación
 - Análisis LPC, obtención de los coeficientes
 - Cálculo del cepstrum a partir de la LPC
 - Análisis de los parámetros obtenidos