Procesamiento Digital de Señales de Voz

Transparencias:

Procesamiento de Señales y Métodos de Análisis para reconocimiento de Voz

Autor: Dr. Juan Carlos Gómez

Basado en: Rabiner, L. and Juang, B-H.. *Fundamentals of Speech Recognition,* Prentice Hall, N.J., 1993.

Introducción

Como ya vimos, los Sistemas de Reconocimiento de Voz comprenden diferentes disciplinas :

- Reconocimiento de patrones Estadísticos
- Teoría de las Comunicaciones
- Procesamiento de Señales
- Matemática Combinatoria
- Lingüística

El denominador común de todo sistema de reconocimiento de voz es la etapa inicial (***front-end***) de procesamiento de señales, que convierte la señal de voz en alguna representación paramétrica para su posterior análisis y procesamiento.

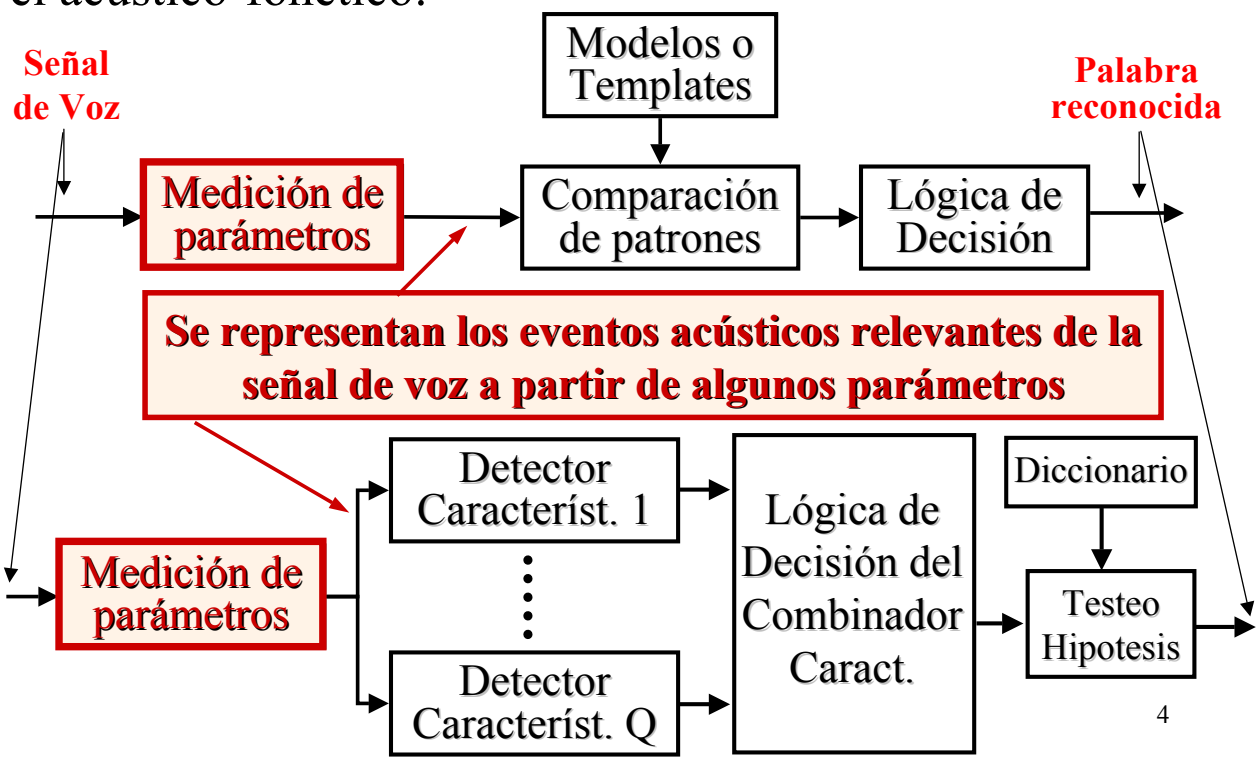
Existe una amplia gama de posibilidades para la representación paramétrica de señales:

- Energía en tiempo corto (*short-time energy*)
- Tasa de cruce por cero (*zero-crossing rate*)
- Tasa de cruce por nivel (*level-crossing rate*)
- Envolvente del espectro en tiempo corto (*short-time spectral envelope*)

Está última forma de representación es la más importante por su difundido uso; y los métodos de análisis espectral son considerados como el núcleo del procesamiento de señales

Modelos de Análisis Espectral

Recordando el enfoque de reconocimiento de patrones y el acústico-fonético:

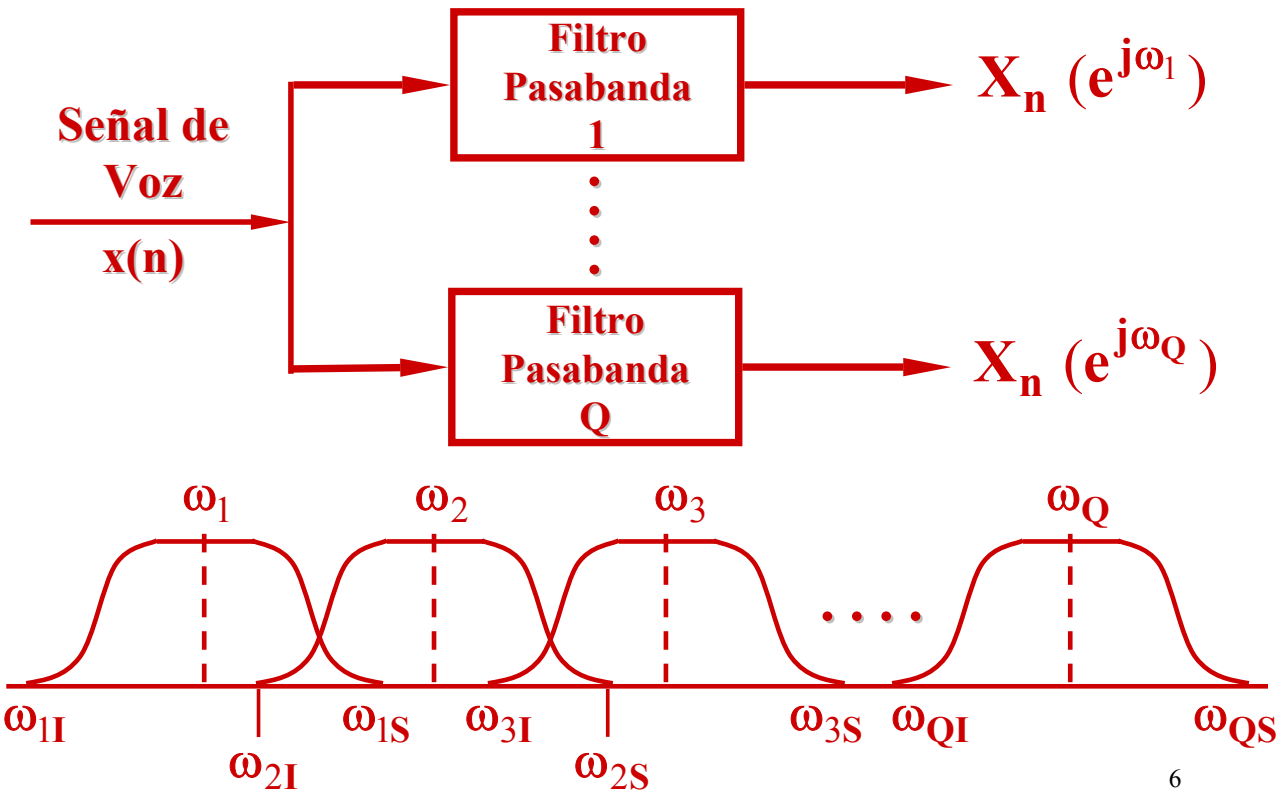


Un buen entendimiento de la forma en la cuál se utilizan las técnicas de procesamiento de señales para implementar la fase de cálculo de parámetros es fundamental para entender los diferentes enfoques para el reconocimiento de voz.

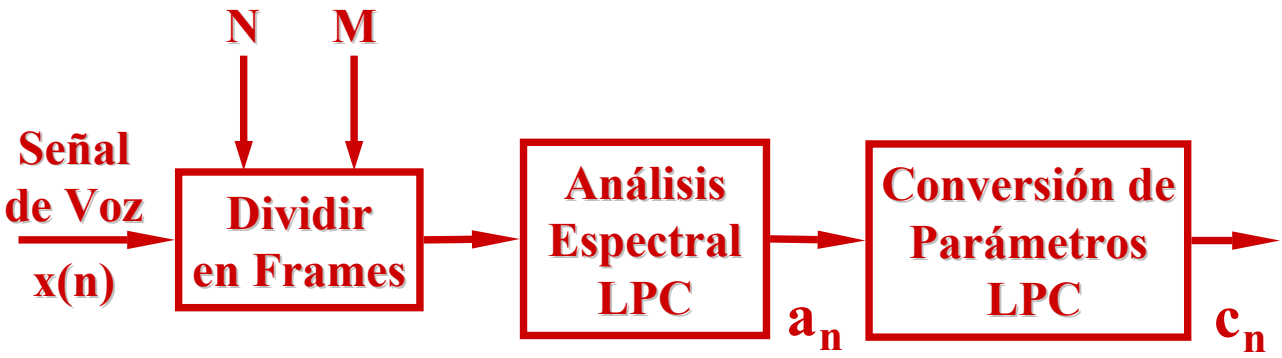
Las 2 técnicas para el procesamiento de señales más comúnmente usadas para reconocimiento de voz son:

- Modelo de Banco de Filtros
- Modelo LPC

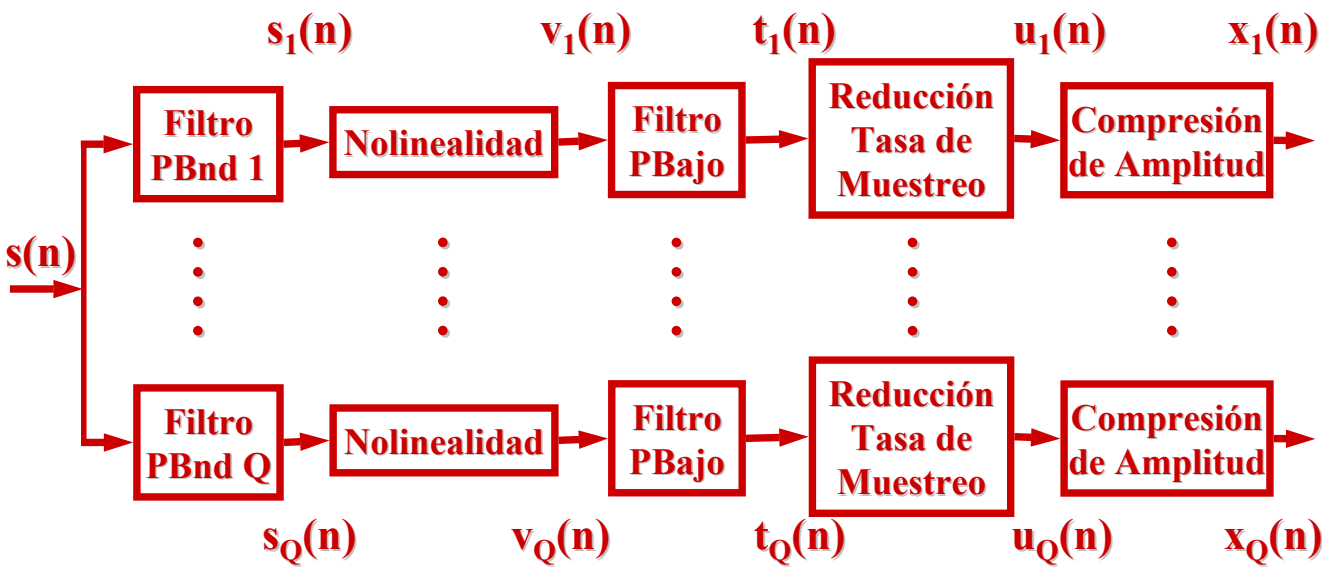
Modelo de Banco de Filtros



Modelo LPC



Etapa Inicial (*Front-End*) de Procesamiento del Banco de Filtros



$$s_i(n) = s(n) * h_i(n) = \sum_{k=0}^{M-1} h_i(k)\, s(n-k) \qquad 1 \le i \le Q$$

$s_i(n)$ se pasa luego por un rectificador de onda completa (o de 1/2 onda). Esto desdobla el espectro de la señal pasabanda a un espectro en la banda de baja frecuencia y otra en la banda de alta frecuencia, que luego se elimina mediante un filtro pasabajos, obteniéndose $u_i(n)$ que representa una estimación de la energía de la señal de voz en cada una de la Q bandas de frecuencia.

Suponiendo que la salida de i-ésimo filtro es una señal del tipo senoidal pura:

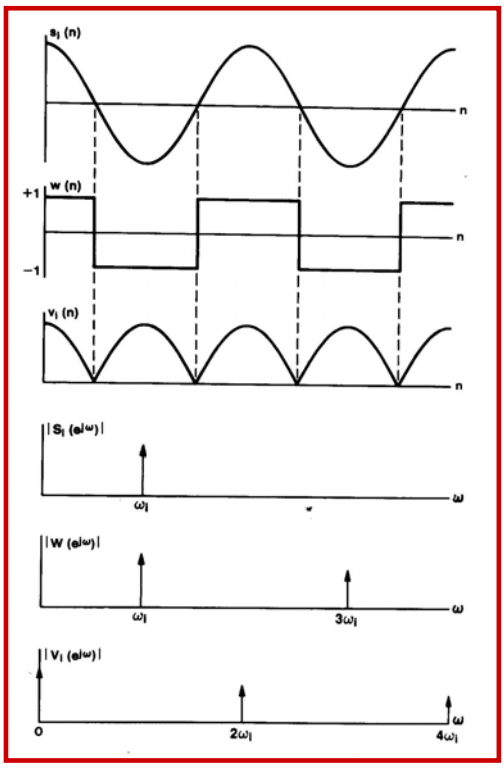
$$s_i(n) = \alpha_i.sen(\omega_i.n)$$

El ancho de banda del filtro es lo suficientemente angosto para dejar pasar un único armónico. Así:

$$v_i(n) = s_i(n).w(n)$$

$$V_i(\omega) = S_i(\omega) * W(\omega)$$

Formas de ondas y espectros típicos para el análisis de una señal senoidal pura

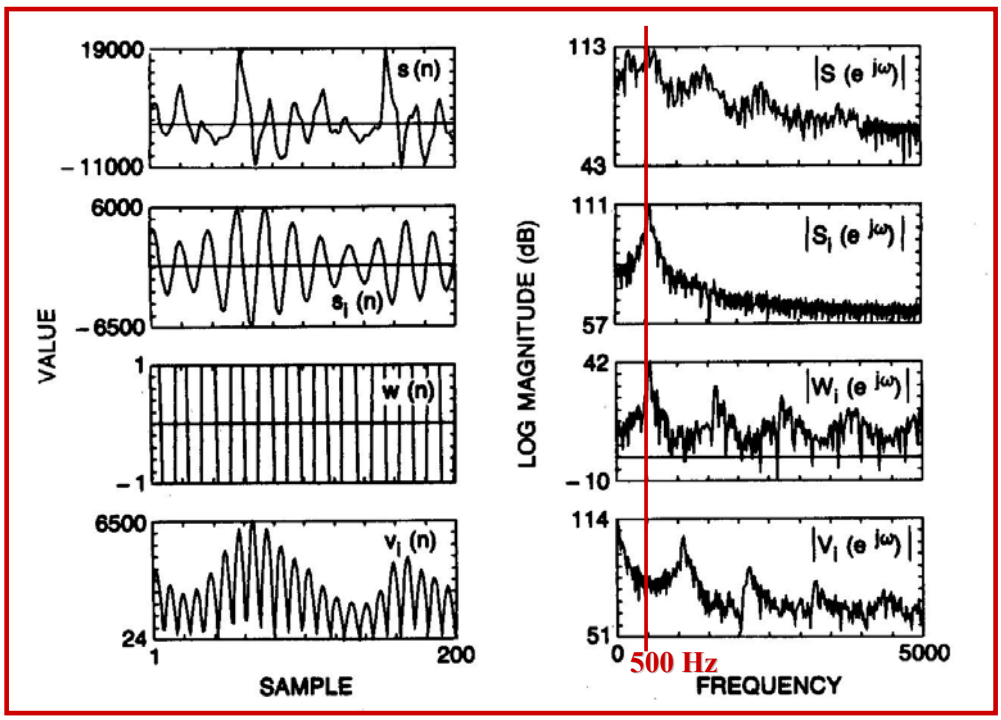


A pesar de que el análisis anterior es estrictamente correcto para señales senoidales puras, es un modelo razonablemente válido para sonidos tonales cuasi-periódicos siempre que el filtro pasabanda no deje pasar 2 o más armónicos de la señal.

La naturaleza de variación en el tiempo de las señales de voz (cuasi-periodicidad) hace que el espectro de la señal ubicado en la banda de baja frecuencia no sea un impulso puro, sino que la información este contenida en una banda.

A continuación se muestra una sección de 20mseg de voz tonal procesada por un canal angosto centrado en 500Hz (y con una F_s=10KHz).

Formas de ondas y espectros de señal de voz con modelo de análisis banco de filtros



El espectro de la señal resultante $\left|V_i\left(e^{j\omega}\right)\right|$ muestra la deseada concentración de energía en baja frecuencia, así como picos no deseados en 1000 Hz, 2000 Hz, etc. La función del filtro Pasa Bajo es eliminar esos picos indeseados. El ancho de banda de la señal $v_i(n)$ está relacionado con la velocidad más rápida de movimiento de los harmónicos de la señal y está en el orden de 20-30 Hz. Los últimos dos bloques en el modelo de banco de filtros son:

• un bloque de reducción de la tasa de muestreo de las señales filtradas con el filtro Pasa-Bajo, $t_i(n)$, que son re-muestreadas con una frecuencia del orden de 40-60 Hz (para obtener una representación más económica), y

• un bloque de compresión del rango dinámico de la señal usando algún esquema de compresión como ser codificación logarítmica, o codificación con ley μ.

Ejemplo: Consideremos un banco de filtros con *Q=16* canales, para una señal de voz de banda ancha con máxima frecuencia de interés de 8 KHz. Se asume una frecuencia de muestreo de $F_s = 20\ KHz$, para evitar aliasing. La tasa de información de la señal sin procesar es

$$20000\frac{\text{muestras}}{\text{seg}}\times 12\frac{\text{bits}}{\text{muestra}} = 240\frac{\text{Kbits}}{\text{seg}}$$

A la salida del analizador, si se usa una frecuencia de muestreo de 50 Hz y una compresión de amplitud logarítmica de 7 bits, se tiene una tasa de información de

$$16\,\text{canales}\times 50\frac{\text{muestras}}{\text{seg.canal}}\times 7\frac{\text{bits}}{\text{muestra}} = 5600\frac{\text{bits}}{\text{seg}}$$

lo que representa una reducción de 40-1 de la tasa de bits.

Tipos de Bancos de Filtros

A) Uniforme: Es el más común, la frecuencia central f_i del i-ésimo filtro pasabanda se define como:

$$f_i = \frac{F_S}{N} i \qquad 1 \le i \le Q$$

donde F_S es la frecuencia de muestreo y N es el número de filtros uniformemente equiespaciados para cubrir el rango de frecuencias de voz. El número de filtros Q satisface la relación:

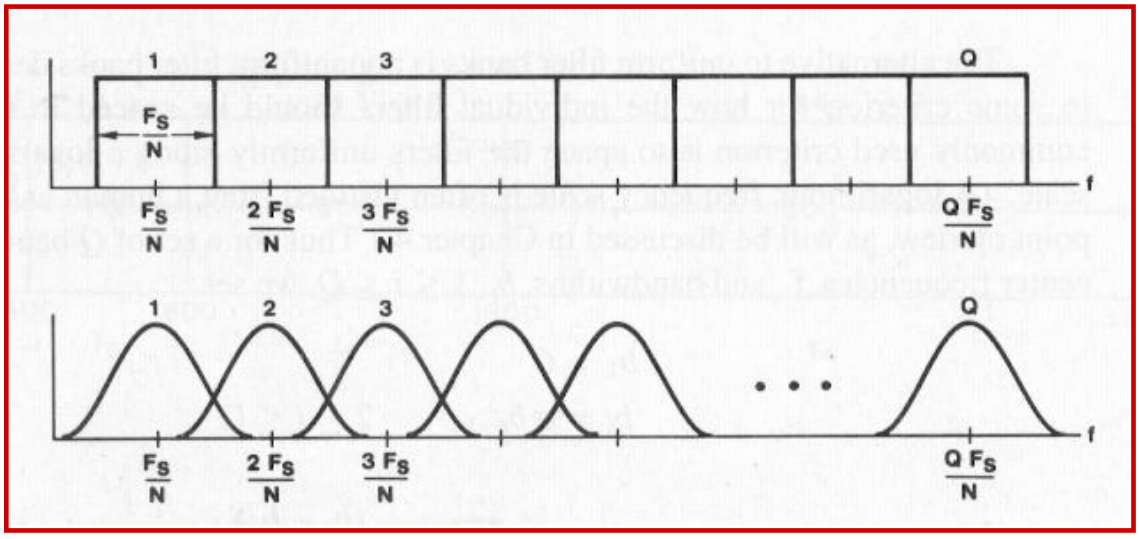
$$Q \le N/2$$

El ancho de banda b_i del i-ésimo filtro satisface:

$$b_i \ge \frac{F_S}{N}$$

la condición de igualdad se verifica si no hay solapamiento

Respuesta ideal y real de un banco de filtros con Q canales



Rango de frecuencias: $F_S/N \longleftrightarrow (Q + 1/2)\, F_S/N$

B) No Uniforme: Se diseñan según algún criterio particular de espaciamiento en frecuencia. Un criterio común es distribuir las frecuencias en forma logarítmica.

Así, para Q filtros pasabanda, la frecuencia central f_i y el ancho de banda b_i del i-ésimo filtro se define como:

$$b_1 = C$$

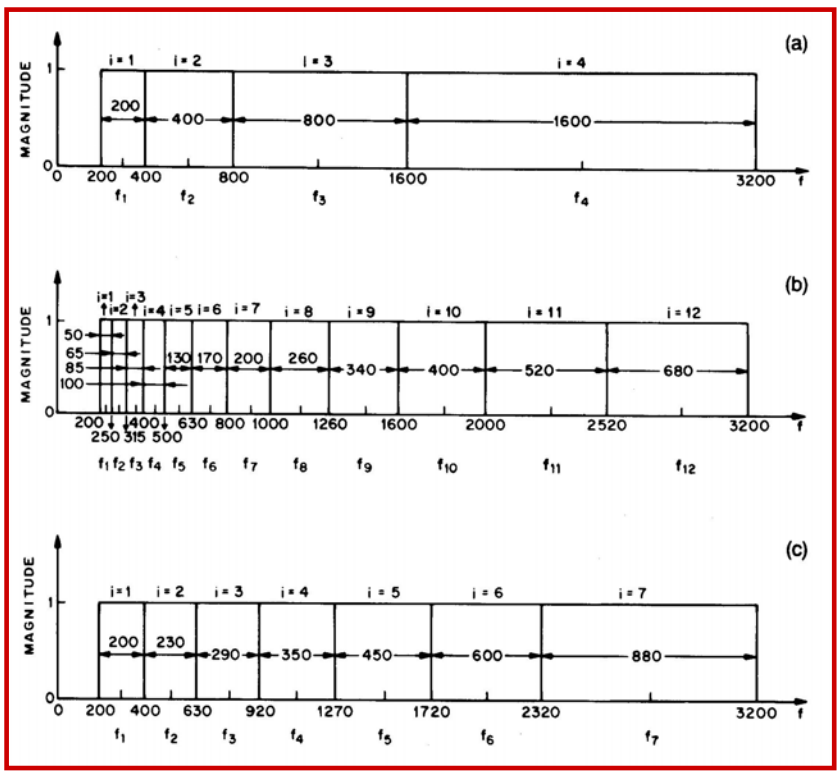
$$b_i = \alpha \, . \, b_{i-1} \qquad 2 \le i \le Q$$

$$f_i = f_1 + \sum_{j=1}^{i-1} b_j + \frac{(b_i - b_j)}{2}$$

donde C y f_1 se definen arbitrariamente.

Los valores de α usados más frecuentemente son $\alpha=2$ que determina un espaciamiento de una octava y $\alpha=4/3$ que determina un espaciamiento de 1/3 de octava.

Especificaciones ideales de bancos de filtros en el rango telefónico (200-3200 Hz)

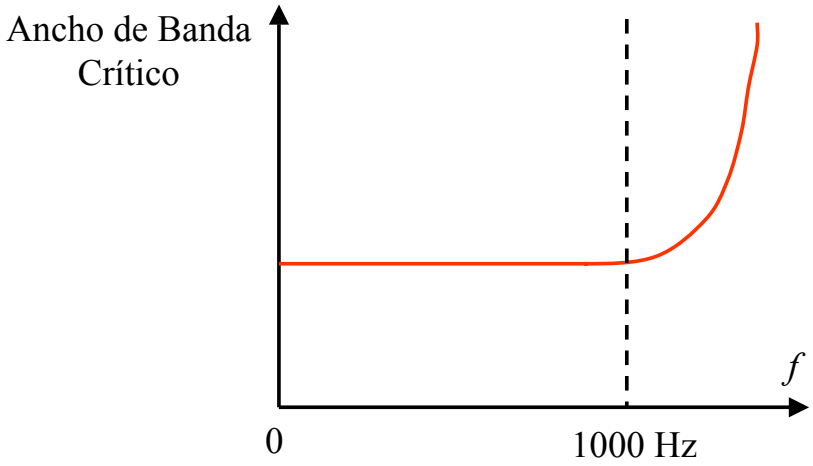


4 canales
1 octava
f_1=300 Hz,
C=200 Hz

12 canales
3 octavas
f_1=225 Hz
C=50 Hz

7 canales
Escala crítica

La **escala crítica** está basada en estudios de percepción del oído humano. La escala es prácticamente lineal hasta frecuencias de aproximadamente 1000 Hz, y para frecuencias superiores es logarítmica (es decir el ancho de banda de los filtros es exponencial en función de la frecuencia).



Implementación de los Bancos de Filtros

Los métodos de diseño de filtros digitales pueden dividirse en 2 clases según el tipo de respuesta al impulso:

1) Filtros IIR (o filtros recursivos): La implementación más eficiente es realizar cada filtro pasabanda individual mediante una estructura cascada o paralelo.

2) Filtros FIR: La implementación más simple es la estructura directa, para la cuál si $h_i(n)$ es la FIR del i-ésimo filtro con L muestras, la salida del filtro es:

$$x_i(n) = s(n) * h_i(n) = \sum_{k=0}^{L-1} h_i(k)\, s(n-k) \qquad 1 \le i \le Q$$

Esta ecuación si bien es simple requiere un alto nivel de requerimiento de cálculos (memoria).

2) <u>Filtros FIR:</u> Una alternativa de implementación menos costosa se da en el caso en que cada respuesta al impulso de los filtros pasabanda puede representarse como una ventana fija pasabajos, *w(n),* modulada por $e^{j\omega n}$, así:

$$h_i(n) = w(n)e^{j\omega_i n}$$

$$x_i(n) = \sum_{k=0}^{L-1} w(k)e^{j\omega_i k}\, s(n-k) = \sum_{k=0}^{L-1} s(k)w(n-k)e^{j\omega_i (n-k)}$$

$$= e^{j\omega_i n}\sum_{k=0}^{L-1} s(k)w(n-k)e^{-j\omega_i k} = e^{j\omega_i n}\, S_n(\omega)$$

donde $S_n(\omega)$ es la transformada de Fourier a corto plazo (Short-Time Fourier Transform) de *s(n)*

Interpretación en el dominio frecuencial de la Transformada de Fourier a Corto Plazo

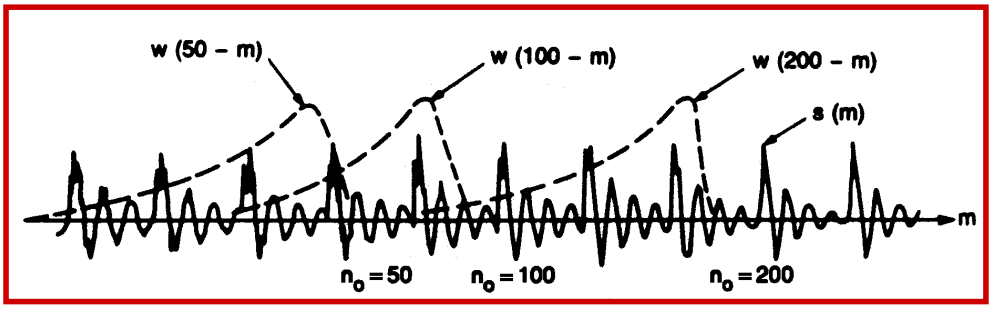
La Transformada de Fourier a Corto Plazo (*STFT*) de la señal *s(n)* se define como:

$$S_n(\omega) = \sum_{k=0}^{L-1} s(k)w(n-k)e^{-j\omega_i k}$$

Para un *n* fijo $n=n_0$

$$S_{n0}(\omega) = \sum_{k=0}^{L-1} s(k)w(n_0-k)e^{-j\omega_i k}$$

obtenemos la FT convencional de la señal truncada, $s(k)w(n_0-k)$, evaluada en la frecuencia $\omega=\omega_i$. A continuación se muestran las señales *s(m)* y $w(n_0-m)$ para n_0= 50, 100 y 200

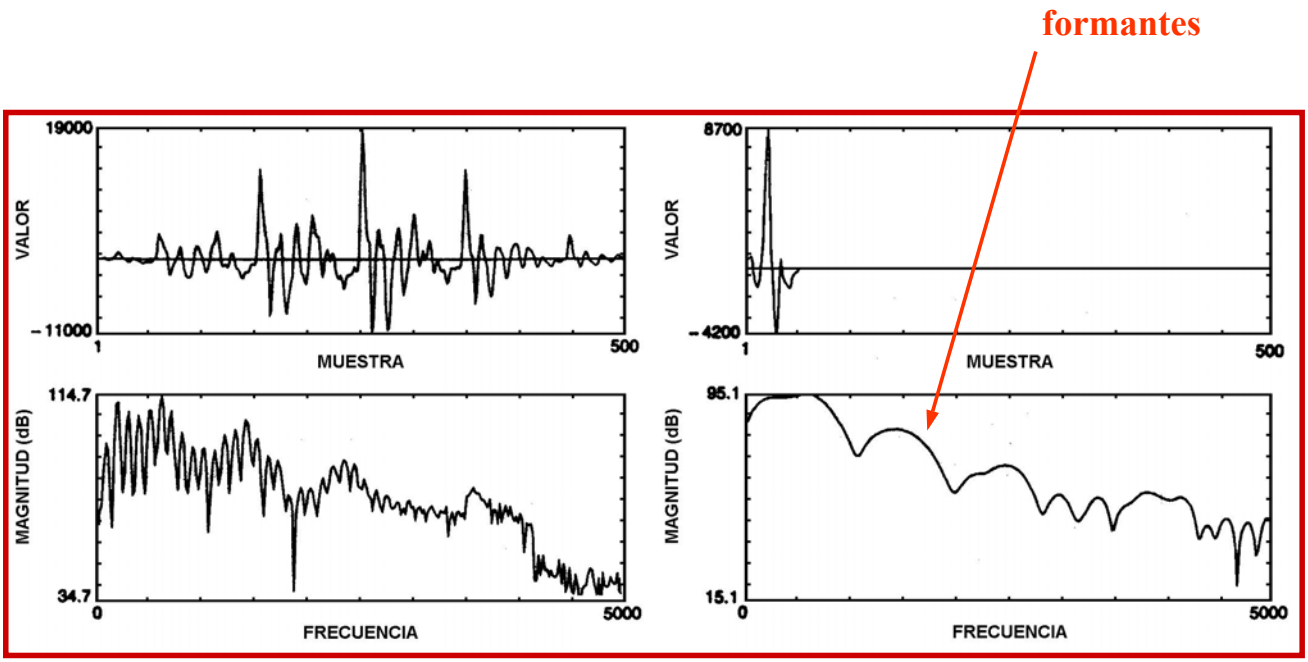


Dado que *w(n)* es un filtro FIR, siendo *L* el número de muestras, podemos establecer que:

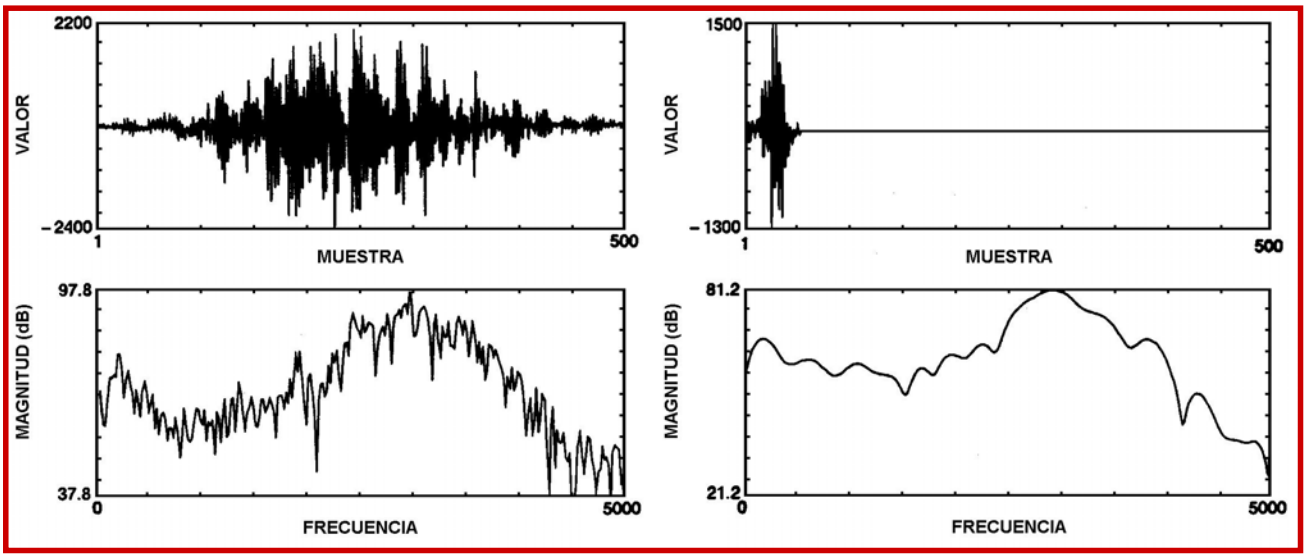
1. Si *L* es grande, comparado con la periodicidad de la señal (pitch), entonces $S_n(\omega)$ tiene buena resolución en frecuencia (podemos visualizar pitchs armónicos individuales) pero sólo se ve a grandes rasgos la envolvente del espectro total en la sección de voz cubierta por la ventana.

2. Si *L* es chico, comparado con la periodicidad de la señal, $S_n(\omega)$ tiene una resolución en frecuencia pobre, pero provee una buena estima de la envolvente del espectro total.

STFT (de 500 y 50 muestras) de una Señal Tonal usando Ventana de Hamming



STFT (de 500 y 50 muestras) de una Señal No Tonal usando Ventana de Hamming

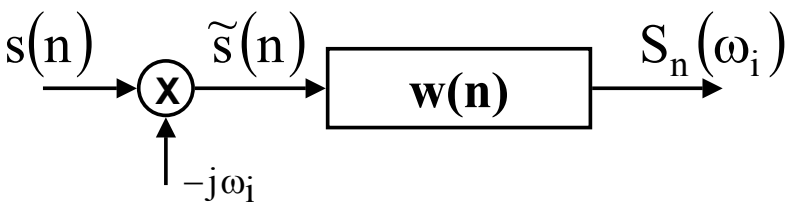


Interpretación de Filtrado Lineal de la Transformada de Fourier a Corto Plazo

La Transformada de Fourier a Corto Plazo (*STFT*) para valores fijos de ω_i queda expresada como:

$$S_n(\omega_i) = s(n)e^{-j\omega_i n} * w(n)$$

siendo esta convolución de tipo circular. Así, podríamos interpretar la fórmula anterior a partir del siguiente gráfico:



$$\text{Como } \tilde{S}(\omega) = S(\omega + \omega_i) \quad \Rightarrow \quad S_n(\omega_i) = S(\omega + \omega_i).W(\omega)$$

Para una ω_i fija la STFT da una representación del espectro de la señal en una banda alrededor de ω_i.

Implementación con FFT del Banco de Filtros Uniforme basado en la STFT

Considerando un banco de filtros uniformemente equiespaciados:

$$f_i = \frac{F_S}{N} i \qquad 1 \le i \le Q$$

Luego:

$$x_i(n) = e^{j\frac{2\pi}{N} i n} \sum_{m=-\infty}^{\infty} s(m) w(n-m) e^{-j\frac{2\pi}{N} i m}$$

Dividiendo la sumatoria en una doble sumatoria de r y k en la cuál:

$$m = N r + k \quad \Leftrightarrow \quad \begin{matrix} 0 \le k \le N-1 \\ -\infty \le r \le \infty \end{matrix}$$

y llamando: $s_n(m)=s(m).w(n-m)$

Obteniéndose entonces:

$$x_i(n) = e^{j\frac{2\pi}{N} i n} \sum_{k=0}^{N-1} \sum_{r=-\infty}^{\infty} s_n(N r + k) e^{-j\frac{2\pi}{N} i (N r + k)}$$

$u_n(k)$

$$= e^{j\frac{2\pi}{N} i n} \sum_{k=0}^{N-1} \left[\sum_{r=-\infty}^{\infty} s_n(N r + k) \right] e^{-j\frac{2\pi}{N} i k}$$

$$= e^{j\frac{2\pi}{N} i n} \left[\sum_{k=0}^{N-1} u_n(k) e^{-j\frac{2\pi}{N} i k} \right]$$

Donde puede verse que $x_i(n)$ es una versión modulada de la DFT de la secuencia $u_n(k)$

Así, los pasos básicos en el cálculo del banco de filtros uniforme a partir de la FFT son:

1. Obtener la señal truncada $s_n(m)=s(m).w(n-m)$ con $m=n-L+1,...,n$ donde $w(n)$ es causal y tiene FIR de L muestras

2. Formar $u_n(k)= \Sigma\, s_n(N\,r + k)$ con $0 \leq k \leq N-1$. Es decir, separa la señal $s_n(m)$ en porciones de N muestras y sumar estas para obtener una señal de N muestras

3. Calcular la DFT de N puntos de $u_n(k)$

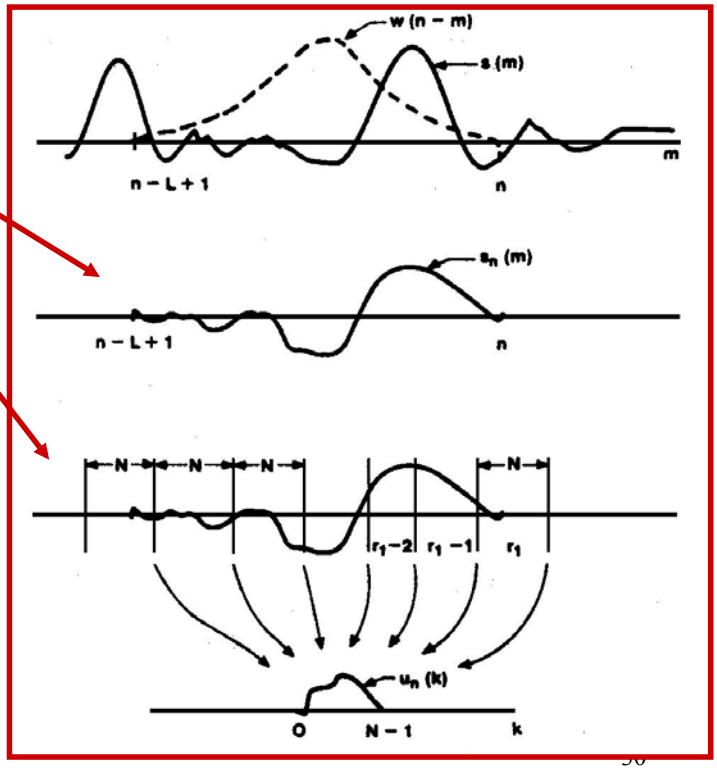
4. Modular la DFT usando la secuencia $e^{j\,2\pi\,i\,n\,/\,N}$ Este paso puede evitarse desplazando en forma circular la secuencia $u_n(k)$ la cantidad $n+N$ muestras para obtener $u_n((k-n))_N$, con $0 \leq k \leq N-1$ antes de calcular la DFT

Pasos básicos en el cálculo del banco de filtros uniforme a partir de la FFT

1. Obtener la señal truncada $s_n(m)=s(m).w(n-m)$

2. Separar la señal $s_n(m)$ en porciones de N muestras y sumar estas para obtener una señal de N muestras

3. Calcular la DFT de N puntos de $u_n(k)$



La cantidad de cálculos para la implementación del banco de filtros es:

$$C_{FBFFT} \cong 2N \log N \qquad \text{multiplicaciones y sumas}$$

Si consideramos a R como la tasa entre las cantidades de cálculos para la implementación directa del banco de filtros y la implementación mediante FFT:

$$R = \frac{C_{DFFIR}}{C_{FBFFT}} = \frac{L\,Q}{2N \log N}$$

Asumiendo N=32 (un banco de filtro de 16 canales) con L=128 (es decir una FIR de 12,8 mseg a una tasa de muestreo de 10KHz) y Q=16 canales, tenemos entonces:

$$R = \frac{128 \,.\, 16}{2 \,.\, 32 \,.\, 5} = 6{,}4$$

La implementación via FFT es 6,4 veces más eficiente.

Implementación del Banco de Filtros no uniforme con FIR

En su forma más general, en un banco de filtros no uniforme cada filtro pasabanda se implementa vía convolución directa, dado que no puede usarse una estructura FFT.

Para el caso en que cada filtro pasabanda se diseña vía el método de ventana, usando la misma ventana pasabajos, podemos demostrar que la respuesta en frecuencia combinada del banco de filtros con Q canales, es independiente del número y la distribución de los filtros individuales.

Para demostrar esto escribimos la respuesta al impulso del *k*-ésimo filtro pasabanda como:

$$h_k(n) = w(n).\tilde{h}_k(n)$$

donde $\tilde{h}_k(n)$ es la respuesta al impulso ideal del filtro pasabanda que se está diseñando. La respuesta en frecuencia puede entonces escribirse como:

$$H_k(\omega) = W(\omega) \circledast \tilde{H}_k(\omega)$$

La respuesta en frecuencia total del banco de filtros es entonces:

$$H(\omega) = \sum_{k=1}^{Q} H_k(\omega) = \sum_{k=1}^{Q} W(\omega) \circledast \tilde{H}_k(\omega) = W(\omega) \circledast \sum_{k=1}^{Q} \tilde{H}_k(\omega)$$

La última sumatoria es la suma de respuestas frecuenciales ideales, donde vemos que es independiente del número y distribución de los filtros individuales y esta sumatoria es 1 en el rango de frecuencias del banco de filtros.

Banco de Filtros No Uniforme basado en FFT

Una forma de aprovechar la estructura de la FFT para crear un banco de filtros no uniformes es mediante la combinación de 2 o 3 canales uniformes. Está técnica equivale a aplicar una ventana modificada a la secuencia antes de aplicarle la FFT.

Para ver esto consideremos la DFT de N puntos de la secuencia *x(n)* derivada de la señal de voz *s(n)* mediante ventaneado con *w(n)*

$$X_k = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}nk} \qquad 0 \le k \le N-1$$

Si sumamos dos DFT de salida tenemos:

$$X_k + X_{k+1} = \sum_{n=0}^{N-1} x(n)(e^{-j\frac{2\pi}{N}nk} + e^{-j\frac{2\pi}{N}n(k+1)})$$

$$= \sum_{n=0}^{N-1} \left[x(n) 2 e^{-j\frac{\pi n}{N}} \cos(\pi n / N)\right] e^{-j\frac{2\pi}{N}nk}$$

La fórmula anterior podría verse como la salida de un canal equivalente donde la secuencia *x(n)* está ponderada en el tiempo por la secuencia compleja .

$$2e^{-j\frac{\pi n}{N}}\cos(\pi n/N)$$

Si se combinan 2 o más canales se puede obtener una secuencia diferente que pondere a la señal. Así combinando canales FFT se pueden obtener rápidamente filtros pasabanda más anchos.
Está técnica constituye un método simple y efectivo para realizar ciertos estructuras de bancos de filtros no uniformes.

Ejemplos Prácticos de Bancos de Filtros para Reconocimiento de Voz

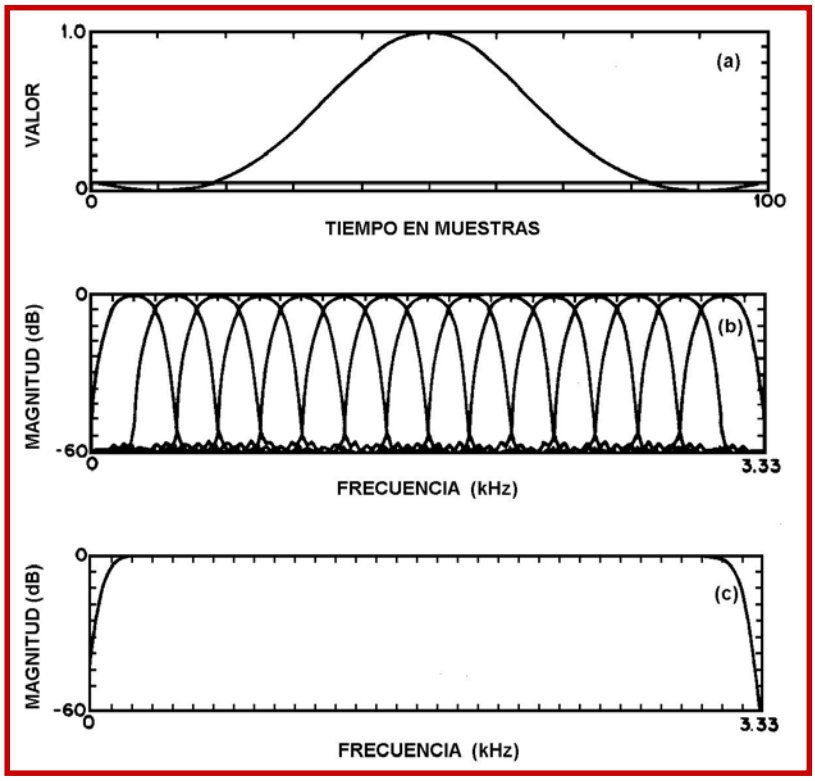
Banco de Filtros uniforme de 15 canales

El filtro básico pasabajos fue diseñado mediante la técnica de ventana usando una de Kaiser de 101 puntos.

En a) se ve la IR del filtro ideal pasabajos multiplicado por la ventana de Kaiser.

En b) se ve la IR de los filtros individuales en la banda de filtrado.

En c) se ve la respuesta en frecuencia total.

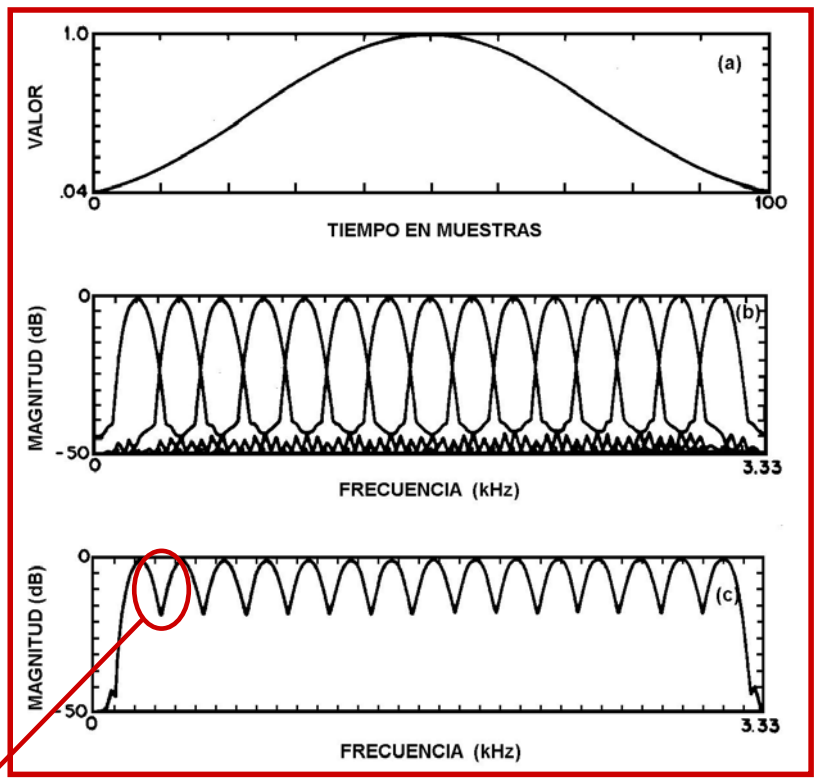


Banco de Filtros uniforme de 15 canales

El filtro pasabajos fue diseñado mediante la técnica de ventana usando una ventana de Kaiser como filtro pasabajos (figura a))

En b) y c) vemos la respuesta en frecuencia de los filtros individuales (más angostos) y del banco total.

Claramente este banco no es aceptable para reconocimiento de voz.



Banco de Filtros No Uniforme de 4 canales

El filtro de una octava cubre el rango de 200 a 3200Hz y fue diseñado usando filtros de fase lineal con FIR.

Cada filtro individual tiene 101 muestras