# Reconocimiento Automático de Voz basado en Técnicas de Comparación de Patrones

**Juan Carlos Gómez**

Presentación basada en las siguientes **Referencias**:

[1] Rabiner, L. & Juang, B-H.. *Fundamentals of Speech Recognition,* Prentice Hall, N.J., 1993.

[2] Rabiner, L. & Juang, B-H.. *Speech Recognition by Machine*, Chap. 47 in *The Digital Signal Processing Handbook*, CRC Press, IEEE Press, 1998.

## Reconocimiento basado en Comparación de Patrones

La característica principal de este enfoque es que usa un marco matemático bien definido y que establece representaciones consistentes de los patrones de voz que pueden usarse para comparaciones confiables a partir de un conjunto de muestras rotuladas, usando algoritmos de entrenamiento. La representación de los patrones de voz puede ser una **plantilla** (template), o un **modelo estadístico** (HMM: Hidden Markov Model), que puede aplicarse a un sonido (más pequeño que una palabra), una palabra, o una frase.

La técnica de reconocimiento consiste básicamente en dos pasos:

- **Primer Paso:** entrenamiento de patrones
- **Segundo Paso:** comparación de patrones

El esquema general está representado en Fig. 1

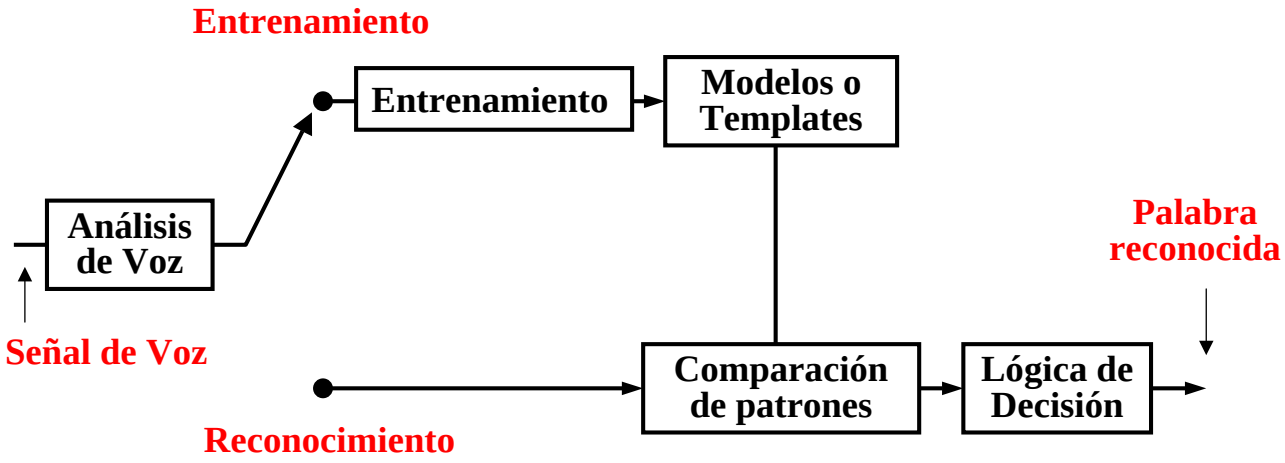


**Fig. 1: Reconocimiento de Voz basado en Comparación de Patrones.**



## ❑ Entrenamiento de patrones

En esta etapa se construye un **patrón de referencia** asociado a cada palabra (o sub-unidad de palabra) que se quiere reconocer, basándose en los vectores característicos de todas las palabras usadas para el entrenamiento. Hay varias formas en que puede realizarse el entrenamiento.

1. **Entrenamiento casual:** un único patrón de sonido es usado para crear el patrón de referencia correspondiente o un modelo estadístico aproximado.
2. **Entrenamiento robusto:** se utilizan varias versiones de cada palabra o sonido a reconocer (generalmente provenientes de un sólo locutor) para construir un patrón de referencia promedio o un modelo estadístico promedio.
3. **Entrenamiento por clustering**: se utiliza un gran número de versiones de cada palabra o sonido (provenientes de un gran número de locutores) para construir patrones de referencia o modelos estadísticos más confiables.

## ❑ Comparación de patrones

En la etapa de comparación de patrones se realiza una comparación directa entre el vector característico asociado a la señal de voz desconocida (a reconocer) y todos los posibles patrones aprendidos en la etapa de entrenamiento, de manera de determinar el mejor ajuste de acuerdo a algún criterio. Surge la necesidad de definir una **medida de similaridad (distancia)** entre vectores característicos que permita determinar cuál es el patrón de referencia que mejor se ajusta a la señal a reconocer.

En general, debido a que una misma palabra es emitida con diferentes velocidades cada vez que es pronunciada, para poder realizar la comparación es necesaria una **normalización temporal,** previa a la extracción de característica. El enfoque más simple sería una transformación lineal del eje tiempo, pero esto no es realista y en la práctica se recurre a técnicas de programación dinámica (**Dynamic Time Warping**).



## ❑ Medidas de Distancia (ó Distorsión)

- Una característica fundamental de los sistemas de reconocimiento (de palabra o de locutor) con el enfoque de **comparación de patrones** es la forma en que los vectores característicos son combinados y comparados con los patrones de referencia.
- Para poder realizar estas operaciones es necesario definir una **medida de distancia** entre vectores característicos.

**Definición:** Una distancia entre dos vectores $\mathbf{x}$ e $\mathbf{y}$ de un espacio vectorial $X$ es una función a valores reales $d(\mathbf{x}, \mathbf{y})$ sobre el producto Cartesiano $X \times X$, que verifica las propiedades

$$(\text{a})\ 0 \le d(\mathbf{x},\mathbf{y}) < \infty, \quad \forall\ \mathbf{x},\mathbf{y} \in X, \ \text{y}\ \ d(\mathbf{x},\mathbf{y}) = 0 \ \text{si y solo si}\ \ \mathbf{x} = \mathbf{y}$$

$$(\text{b})\ d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x}), \quad \forall\ \mathbf{x},\mathbf{y} \in X$$

$$(\text{c})\ d(\mathbf{x},\mathbf{y}) \le d(\mathbf{x},\mathbf{z}) + d(\mathbf{y},\mathbf{z}), \quad \forall\ \mathbf{x},\mathbf{y},\mathbf{z} \in X$$

- Algunas de las medidas de distancia más utilizadas son las distancias o **métricas inducidas por las normas en espacios** $L_p$.

  Por ejemplo, si $f_i$, $f_i^{'}$ con $i = 1,2,\cdots,D$ son las componentes de dos vectores característicos $f$ y $f^{'}$, pueden definirse las siguientes métricas inducidas por las normas $L_p$

$$d_1(f,f') = \sum_{i=1}^{D} \left| f_i - f_i^{'} \right|$$ **distancia** $L_1$

$$d_2(f,f') = \sqrt{\sum_{i=1}^{D} \left( f_i - f_i^{'} \right)^2}$$ **distancia Euclidea (o** $L_2$ **)**



- Una medida de distancia muy utilizada cuando se emplean como característica los coeficientes cepstral, y que ha probado tener una muy buena performance en tareas de reconocimiento, es la **distancia Euclidea ponderada**, definida como

$$d_{2w}(c,c') = \sqrt{\sum_{i=1}^{D} \left( w_i \left( c_i - c_i^{'} \right) \right)^2}$$

donde

$$w_i = \frac{1}{\sigma_i}$$

siendo $\sigma_i^2$ una estima de la varianza del $i$-ésimo coeficiente cepstral $c_i$. Aquí, los datos que son menos confiables (con mayor varianza o dispersión) son pesados menos.

- Cuando se utiliza el **power cepstrum** como vector característico, a la distancia en $L_2$ se la denomina **distancia cepstral.** Teniendo en cuenta la identidad de Parseval, resulta

$$d_{\text{ceps}}(c,c') = \sqrt{\sum_{m=-\infty}^{\infty}\left(c_m - c_m^{'}\right)^2} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\ln\left(\left|X(\omega)\right|^2\right) - \ln\left(\left|X'(\omega)\right|^2\right)\right|^2 d\omega}$$

En la práctica, sólo se computa un número finito $Q$ de coeficientes **power cepstral**, resultando

$$d_{\text{ceps}}(c,c') \cong \sqrt{\sum_{m=1}^{Q}\left(c_m - c_m^{'}\right)^2}$$



## • Distancia de Mahalanobis

Cuando el vector característico no es homogéneo, en el sentido de que está compuesto por características de distinta naturaleza, puede suceder que las distintas componentes estén en distintas escalas. Al usar la norma Euclidea, o la norma 1, para calcular la distancia entre vectores, las componentes en las escalas mas pequeñas casi no tendrán influencia en el valor de la distancia, a pesar de que esas componentes puedan diferir mucho. Surge entonces la necesidad de usar distancias normalizadas. Una posible forma de normalizar la distancia entre un vector $x$ y la media $\mu$ del conjunto de vectores al cual pertenece $x$, para el caso escalar, es definiendo

$$r = \left|\frac{x-\mu}{\sigma}\right| \qquad (1)$$

donde $\sigma$ es el desvio estándar del conjunto de vectores, y $r$ es la distancia normalizada.

De esta forma un vector que está separado un desvio standard de la media, estará a una distancia normalizada igual a 1 (independientemente de las unidades en que se mida la distancia).

Esta idea se puede generalizar para el caso de un vector de dimensión *d,* definiendo

$$r^2 = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 + \cdots + \left(\frac{x_d - \mu_d}{\sigma_d}\right)^2$$

Esta definición tiene sentido si cada componente del vector característico mide propiedades independientes del patrón a reconocer.

Considerando la expresión (1) para el caso escalar, la distancia normalizada puede escribirse



$$r^2 = (x - \mu)^T \frac{1}{\sigma^2}(x - \mu)$$

donde $\sigma^2$ es la varianza del conjunto de vectores, que para el caso multidimensional resulta

$$r^2 = (x - \mu)^T \Sigma^{-1}(x - \mu)$$

donde $\Sigma$ es la matriz de covarianza del conjunto de vectores.

Esta formulación de la distancia es más general ya que tiene en cuenta la interacción entre coeficientes a traves de la matriz de covarianza y se la denominada **Distancia de Mahalanobis.**

Si se tienen $N_j$ vectores de entrenamiento para cada una de las $M$ palabras del vocabulario, la **distancia de Mahalanobis** entre un vector a reconocer $x$, y el vector patrón representativo de la palabra $j$, resulta

$$d_M(x,\mu_j)=\sqrt{(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)}$$

donde

$$\mu_j=\frac{1}{N_j}\sum_{k=1}^{N_j}x_{kj}^e$$

Media (muestral) de los vectores de entrenamiento $x_{kj}^e$

$$\Sigma_j=\frac{1}{N_j-1}\sum_{k=1}^{N_j}(x_{kj}^e-\mu_j)(x_{kj}^e-\mu_j)^T$$

Matriz de Covarianza (muestral) de los vectores de entrenamiento $x_{kj}^e$

$x$ : vector característico asociado a la palabra a reconocer



## ❑ Clasificador de Mínima Distancia

Denotemos con $\mu_1, \mu_2, \ldots, \mu_M$ a los vectores promedio representativos de los patrones 1, 2, ..., $M$, respectivamente, obtenidos computando la media muestral de los datos de entrenamiento correspondientes a cada palabra en el vocabulario. Es decir

$$\mu_j=\frac{1}{N_j}\sum_{k=1}^{N_j}x_{kj}^e \qquad j=1, 2, \ldots, M$$

Donde $\{x_{kj}^e\}_{k=1}^{N_j}$ es el conjunto de datos de entrenamiento de la palabra $j$ del vocabulario.

La Fig. 2, representa los datos de entrenamiento agrupados en clusters para el caso de vectores característicos de dimensión 3.

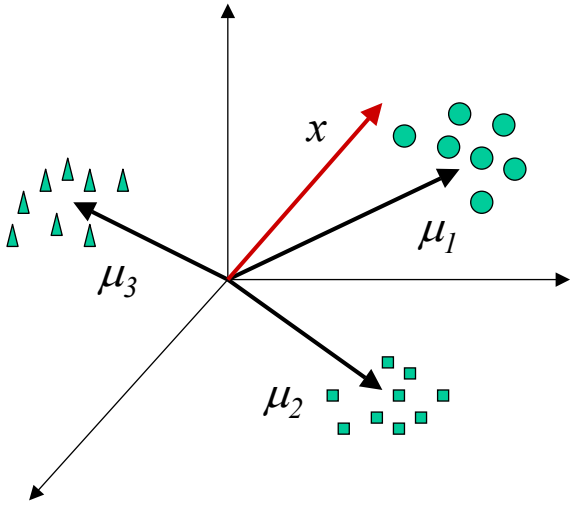


**Fig. 2: Datos de entrenamiento agrupados en clusters correspondientes a las distintas clases o palabras del vocabulario.**

Una manera obvia de clasificar (reconocer) una nueva palabra *x* es computar la distancia de *x* a cada uno de los patrones $\mu_1$, $\mu_2$ , ..., $\mu_M$ y asignar a *x* la palabra del vocabulario cuyo patrón está a la mínima distancia. Por este motivo este clasificador (reconocedor) se denomina de **mínima distancia.**



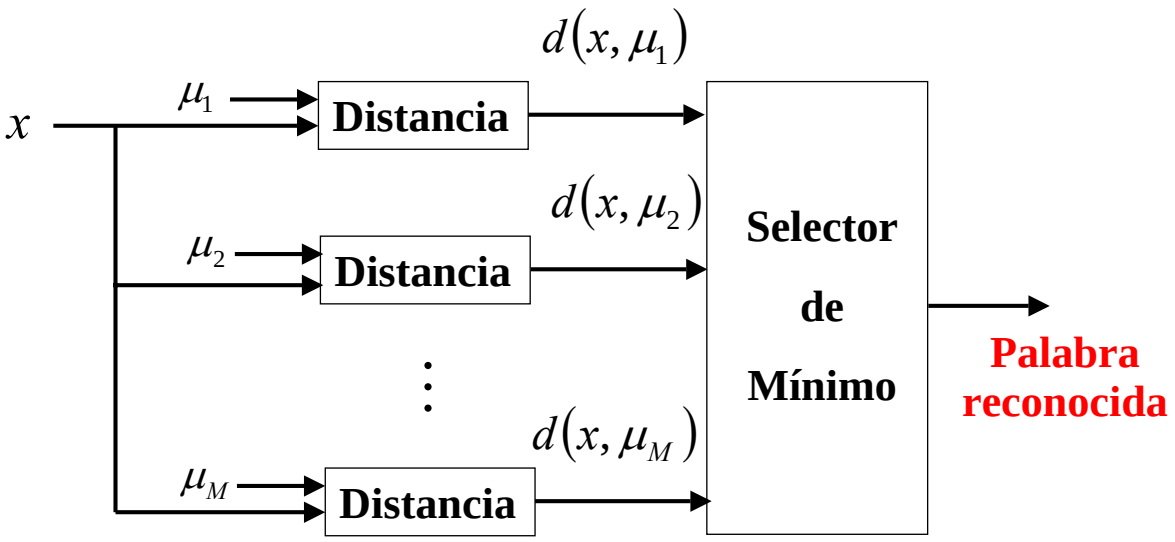


**Figura 3: Diagrama en bloques de un clasificador de mínima distancia.**

## ❑ Etapa de Análisis de Voz (Extracción de Característica)

La primera etapa en el procesamiento (que es común a todos los enfoques de reconocimiento) es la etapa de **análisis de voz y extracción de característica**. Se computa una representación (espectral) de las características inestacionarias de la señal de voz. Estas medidas espectrales se convierten luego en un conjunto de parámetros que describen la propiedades acústicas de las unidades fonéticas. Estos parámetros pueden ser: nasalidad (presencia o ausencia de resonancia nasal), fricación (presencia o ausencia de excitación aleatoria en la voz), ubicación de los formantes (frecuencias de las 3 primeras resonancias), clasificación entre sonidos tonales y no tonales, **coeficientes cepstral**, energía de la señal, etc.

Los métodos más comunes en esta etapa son **análisis con Banco de Filtros** y **Análisis LPC (Linear Predictive Coding)**.



## ❑ Análisis LPC

El análisis con modelos LPC es el más usado en la primera etapa de procesamiento de los sistemas de reconocimiento de voz, principal-mente debido a que :

- los modelos LPC proveen una buena representación de la señal de voz, especialmente cuando se trata de sonidos tonales.
- los modelos LPC proveen una buena separación entre la excitación y el modelo del tracto vocal, permitiendo una representación simple.
- los modelos LPC son simples, matemáticamente precisos y relativamente fáciles de implementar tanto en software como en hardware. La carga computacional es menor que la que requeriría la implementación de procesamiento usando banco de filtros.
- la experiencia muestra que los modelos LPC funcionan bien en tareas de reconocimiento.

Las tareas de procesamiento involucradas en el Análisis LPC se representan en la Fig. 2.

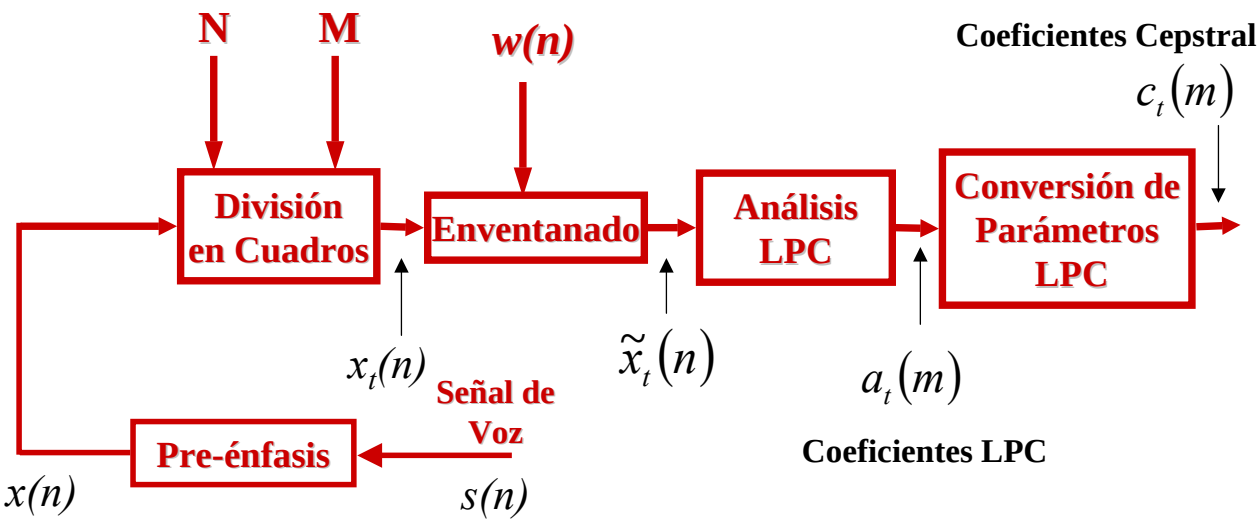


**Fig. 2: Análisis LPC**



## Pre-énfasis

Para hacer menos sensible al sistema a los efectos de cuantización por longitud finita de palabra, se pasa a la señal de entrada por un filtro de bajo orden (típicamente un filtro FIR de primer orden) de manera de ***aplanar*** su espectro.

Un filtro de pre-énfasis típico es

$$H(z) = 1 - az^{-1}, \qquad 0.9 \le a \le 1.0$$

Un valor muy usado de $a$ es 0.95 (para implementaciones en punto fijo se suele utilizar $a$ = 15/16 = 0.9375).

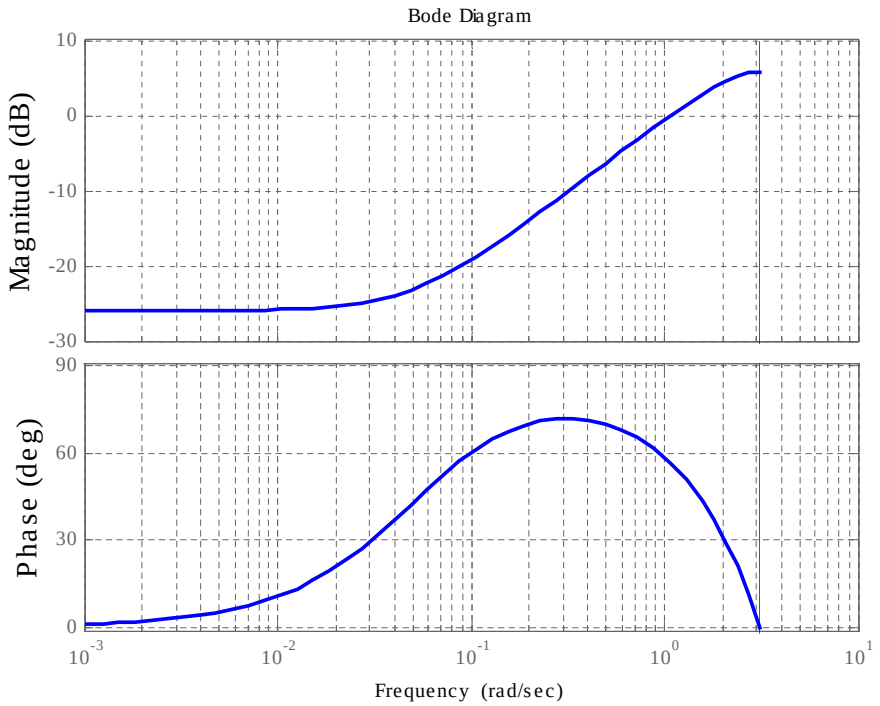


**Fig. 3: Respuesta en frecuencia del filtro de pre-énfasis**



## División en cuadros

La señal de voz es dividida en cuadros (***frames***) de duración entre 10 y 30 mseg donde se asume que la señal es cuasi-estacionaria. El número de muestras por cuadro es $N$, y cuadros adyacentes están separados M muestras. Existe entonces un **solapamiento** de $N - M$ muestras entre cuadros adyacentes. Se logra con esto que los coeficientes LPC tengan una transición suave entre un cuadro y el siguiente. La Fig. 4 esquematiza este procesamiento.

Valores típicos para $N$ y $M$ son 240 y 80 para una frecuencia de muestreo de 8 KHz. Esto resulta en cuadros de 30 mseg. de duración, separados 10 mseg., lo que equivale a una tasa de cuadro de 100 Hz.

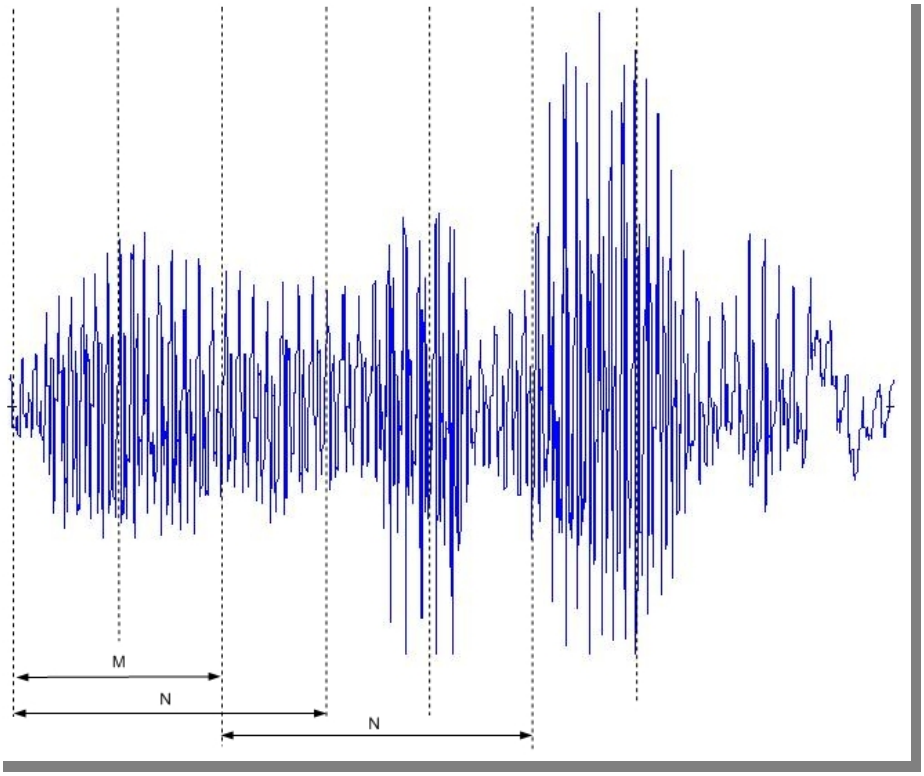


**Fig. 4: División en cuadros**



## Enventanado

Cada cuadro es pasado por un ventana para minimizar las discontinuidades de la señal al principio y al final de cada cuadro. La señal resultante es

$$\tilde{x}_t(n) = x_t(n)w(n), \qquad 0 \le n \le N-1$$

Una ventana ampliamente utilizada en sistemas de reconocimiento de voz es la ventana de Hamming.

$$w(n) = 0.54 - 0.46\ \cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \le n \le N-1$$

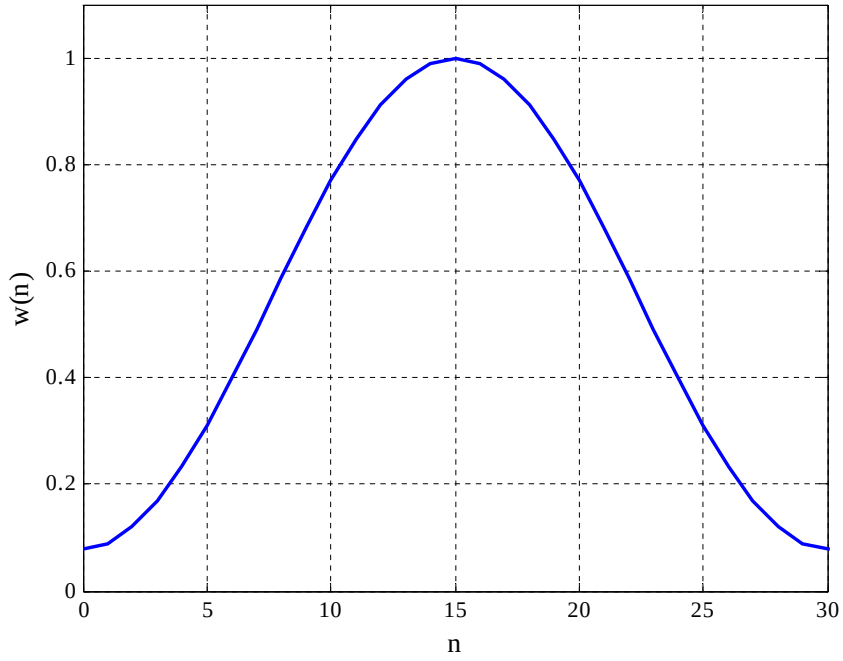


**Fig. 5: Ventana de Hamming**



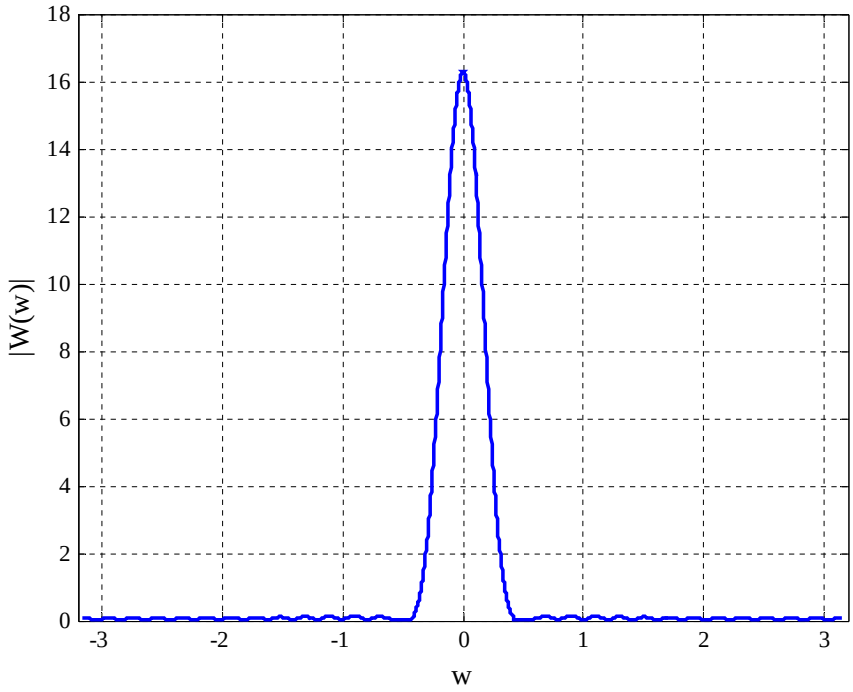


**Fig. 6: Espectro de la Ventana de Hamming**

## Análisis LPC

Para cada cuadro se computan los coeficientes LPC (entre 10 y 20 coeficientes), generalmente basándose en la **autocorrelación** de la señal enventanada. Esto implica una **reducción de la tasa de información**.

Por ejemplo para cuadros de 20 mseg., con frecuencia de muestreo de 8 KHz, y con 8 bits por muestra, resulta una tasa de información de **1280 bits/cuadro**, en tanto que si se consideran 15 coeficientes LPC por cuadro, resultaría una tasa de información de 15 x 8 = **120 bits/cuadro**, lo que representa una **reducción de la tasa de información de 1280/120 = 10.67 veces.**



## Conversión a Coeficientes Cepstral

Sea una secuencia real $\{x(n)\}$ con transformada $Z$ , $X(z)$. Se asume que $X(z)$ converge en la circunferencia unitaria. El **cepstrum complejo** de la señal $\{x(n)\}$ se define como la secuencia $\{c_x(n)\}$ que es la transformada $Z$ inversa de $C_x(z)$ , donde

$$C_x(z) = \ln(X(z))$$

El cepstrum complejo existe si $C_x(z)$ converge en la región anular $r_1 < |z| < r_2$ , donde $0 < r_1 < 1, \quad r_2 > 1$ . Esto implica que $C_x(z)$ converge en la circunferencia unitaria, es decir

$$C_x(\omega) = \ln(X(\omega)) = \sum_{n=-\infty}^{\infty} c_x(n) e^{-j\omega n}$$

y entonces el cepstrum complejo puede obtenerse como la transformada inversa de Fourier de $C_x(\omega) = \ln(X(\omega))$, es decir

$$c_x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(X(\omega)\right) e^{j\omega n} d\omega$$

Si se expresa $X(\omega)$ en términos de su magnitud y fase

$$X(\omega) = |X(\omega)| e^{j\theta(\omega)}$$

entonces

$$\ln\left(X(\omega)\right) = \ln\left(|X(\omega)|\right) + j\theta(\omega)$$

y el cepstrum complejo resulta

$$c_x(n) = \underbrace{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(|X(\omega)|\right) e^{j\omega n}\, d\omega}_{c_m(n)} + \underbrace{j\frac{1}{2\pi}\int_{-\pi}^{\pi} \theta(\omega) e^{j\omega n} d\omega}_{c_\theta(\omega)}$$



A la secuencia $\{c_m(n)\}$ se la denomina **cepstrum real**, o simplemente **cepstrum**, en tanto que a los $c_m(n)$ se los denomina **coeficientes cepstral (reales)**.

En las aplicaciones de reconocimiento de voz los coeficientes cepstral (reales) han probado ser un conjunto de características más robustas y confiables que los coeficientes LPC. En la práctica, suele usarse, en lugar del ceptrum real, el **power cepstrum (real)** definido como

$$c_m^{power}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(|X(\omega)|^2\right) e^{j\omega n}\, d\omega = 2c_m(n)$$

Para confundir un poco las cosas al power cepstrum también se lo suele denominar simplemente **cepstrum**.

Cuando la señal de voz *x(n)* es modelada como la salida de un sistema (impropiamente denominado) **all-pole** (sólo polos) de mínima fase, como el modelo LPC, entonces los coeficientes power cepstral pueden calcularse en forma recursiva a partir de los coeficientes LPC. En este caso se los denomina **coeficientes power cepstral LPC**.

Sea *x(n)* la señal de voz modelada como

$$x(n) = \frac{1}{A\left(q^{-1}\right)} e(n)$$

donde *e(n)* es ruido blanco con media cero y varianza $\sigma^2$, y

$$A\left(q^{-1}\right) = 1 + \sum_{k=1}^{p} a_k q^{-k} = \sum_{k=0}^{p} a_k q^{-k} \quad , \quad a_0 = 1$$

donde $a_k$ , $k = 1, \cdots, p$ , son los coeficientes LPC.



La densidad espectral de energía resulta entonces

$$|X(\omega)|^2 = \frac{\sigma^2}{|A(\omega)|^2}$$

de donde

$$\ln\left(|X(\omega)|^2\right) = \ln \sigma^2 - \ln\left(|A(\omega)|^2\right)$$

$$= \sum_{n=-\infty}^{\infty} c_m^{power}(n) e^{-j\omega n}$$

Por otra parte

$$\ln\left(|A(\omega)|^2\right) = \ln\left(A\left(e^{j\omega}\right)A\left(e^{-j\omega}\right)\right) = \ln\left(\sum_{k=0}^{p} a_k e^{j\omega k} \sum_{n=0}^{p} a_n e^{-j\omega n}\right)$$

En el dominio transformado Z, resulta

$$\ln\left(\left|X(z)\right|^2\right)=\sum_{n=-\infty}^{\infty}c_m^{power}(n)z^{-n}=\ln\sigma^2-\ln\left(\sum_{k=0}^{p}a_k z^k\sum_{n=0}^{p}a_n z^{-n}\right)$$

Derivando ambos miembros respecto a $z$ , e igualando los coeficientes de potencias de $z$ del mismo orden, se obtiene la siguiente ecuación recursiva para los coeficientes power cepstral

$$c_m^{power}(n)=-a_n-\frac{1}{n}\sum_{k=1}^{n-1}k\,c_m^{power}(k)\,a_{n-k}\,,\qquad n>0$$

$$c_m^{power}(0)=\ln\sigma^2$$

con $a_0 = 1$ y $a_k = 0$ para $k > p$ .



Usualmente se computa un número $Q$ de coeficientes cepstral por cuadro, donde $Q > p$ , siendo $p$ el número de coeficientes LPC por cuadro. Se suele adoptar

$$Q\approx\frac{3}{2}p$$

## Valores Típicos de parámetros en Análisis LPC

| Parámetro | $F_s$ = 6.67 KHz | $F_s$ = 8 KHz | $F_s$ = 10 KHz |
|---|---|---|---|
| $N$ | 300 (45 mseg) | 240 (30 mseg) | 300 (30 mseg) |
| $M$ | 100 (15 mseg) | 80 (10 mseg) | 100 (10 mseg) |
| $p$ | 8 | 10 | 10 |
| $Q$ | 12 | 12 | 12 |