

# Lección 3: Sistemas de Reconocimiento de Voz

Dr. Jesús Savage  
Dr. Carlos Rivera

2 de marzo de 2022

# Índice

- 1 Medidas de comparación de señales.
- 2 Procesamiento y reconocimiento de señales de voz utilizando técnicas de LPC.

# Medidas de comparación de señales

Supongase que se cuenta con dos vectores de características  $\underline{x}$  y  $\underline{y}$  definidos en el espacio vectorial  $\chi$ .

Una función de distancia o métrica es una función real tal que:

- 1  $0 \leq d(\underline{x}, \underline{y}) < \infty$  para  $\underline{x}$  y  $\underline{y} \in \chi$   
y  $d(\underline{x}, \underline{y}) = 0$  si y solo si  $\underline{x} = \underline{y}$ .
- 2  $d(\underline{x}, \underline{y}) = d(\underline{y}, \underline{x})$
- 3  $d(\underline{x}, \underline{y}) \leq d(\underline{x}, \underline{z}) + d(\underline{y}, \underline{z})$  para  $\underline{x}, \underline{y}, \underline{z}$ .
- 4  $d(\underline{x} + \underline{z}, \underline{y} + \underline{z}) = d(\underline{x}, \underline{y})$ .

# Medidas de comparación de señales

Para el reconocimiento de voz se compara el espectro de la señal con el modelo del espectro. Si la transformada de Fourier de una señal  $x(t)$  es  $X(j\omega) = F(x(t))$ , su espectro es:  $S(\omega) = |X(j\omega)|^2 = X(j\omega) \cdot X(-j\omega)$

$$= \int_{-\infty}^{\infty} x(t)e^{-j\omega t} dt \cdot \int_{-\infty}^{\infty} x(l)e^{+j\omega l} dl$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(t) \cdot x(l)e^{-j\omega t} e^{+j\omega l} dt dl$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(t) \cdot x(l)e^{-j\omega(t-l)} dt dl$$

## Medidas de comparación de señales

Haciendo:  $t - l = \tau \rightarrow dt = d\tau$

$t = l + \tau$ ; si  $t = -\infty \rightarrow \tau = -\infty$ ;  $t = \infty \rightarrow \tau = \infty$

Por lo tanto:

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(t) \cdot x(l) e^{-j\omega(t-l)} dt dl = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} x(l + \tau) \cdot x(l) dl \right\} e^{-j\omega\tau} d\tau$$

$$= \int_{-\infty}^{\infty} r(\tau) e^{-j\omega\tau} d\tau = S(\omega)$$

Por lo tanto el espectro de  $x(t)$  es la transformada de Fourier de su función de correlaciones  $r(\tau)$ :

$$S(\omega) = |X(j\omega)|^2 = X(j\omega) \cdot X(-j\omega) = F(r(\tau))$$

# Medidas de comparación de señales

Usando el modelo de generación de voz:

$$H(z) = \frac{\sigma}{\sum_{i=0}^M a_i \cdot z^{-i}}$$

con  $z = e^{j\omega}$ , el espectro es:

$$\begin{aligned}\hat{S}(\omega) &= |H(e^{j\omega})|^2 = H(e^{j\omega}) \cdot H(e^{-j\omega}) = \frac{\sigma}{\sum_{i=0}^M a_i e^{-ij\omega}} \cdot \frac{\sigma}{\sum_{i=0}^M a_i \cdot e^{+ij\omega}} \\ &= \frac{\sigma^2}{|A(e^{j\omega})|^2}\end{aligned}$$

# Medidas de comparación de señales

Una medida de comparación de señales de voz está basado en los estimadores de máxima verosimilitud (likelihood), llamada distancia de Itakura Saito:

$$d_{is}(S(\omega), S'(\omega)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} [e^{V(\omega)} - V(\omega) - 1] d\omega$$

$$\text{Con } V(\omega) = \ln S(\omega) - \ln S'(\omega)$$

Recordar que:

$$\log_{10}(1000) = 3 \quad \rightarrow \quad 10^{\log(1000)} = 1000$$

$$\text{y } e^{\ln(n)} = n$$

Por lo tanto:

$$d_{is}(S(\omega), S'(\omega)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} [e^{V(\omega)} - V(\omega) - 1] d\omega = \\ \frac{1}{2\pi} \int_{-\pi}^{\pi} [e^{(\ln S(\omega) - \ln S'(\omega))} - (\ln S(\omega) - \ln S'(\omega)) - 1] d\omega$$

## Medidas de comparación de señales

$$d_{is}(S(\omega), S'(\omega)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ \frac{S(\omega)}{S'(\omega)} - (\ln S(\omega) - \ln S'(\omega)) - 1 \right] d\omega$$

Con  $S'(\omega) = \frac{\sigma^2}{|A(e^{j\omega})|^2}$  el primer termino de la integral es:

$$\begin{aligned} d_{is}(S(\omega), \hat{S}(\omega)) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{S(\omega)}{S'(\omega)} d\omega = \frac{1}{2\pi\sigma^2} \int_{-\pi}^{\pi} S(\omega) |A(e^{j\omega})|^2 d\omega \\ &= \frac{1}{2\pi\sigma^2} \int_{-\pi}^{\pi} S(\omega) \sum_{i=0}^M a_i e^{j\omega i} \cdot \sum_{k=0}^M a_k \cdot e^{-j\omega k} d\omega \\ &= \frac{1}{2\pi\sigma^2} \sum_{i=0}^M a_i \cdot \sum_{k=0}^M a_k \int_{-\pi}^{\pi} S(\omega) \cdot e^{j\omega(i-k)} d\omega \\ &= \frac{1}{\sigma^2} \sum_{i=0}^M a_i \cdot \sum_{k=0}^M a_k r(i-k) \\ &= \frac{1}{\sigma^2} [a_0 a_0 r(0) + a_0 a_1 r(-1) + \dots + a_0 a_M r(-M) + \\ & a_1 a_0 r(1) + a_1 a_1 r(0) + \dots + a_1 a_M r(1-M) + \\ & + \dots + a_M a_0 r(M) + a_M a_1 r(M-1) + \dots + a_M a_M r(0)] = \frac{1}{\sigma^2} [\underline{a}^T R \underline{a}] \end{aligned}$$

## Medidas de comparación de señales

$$\begin{aligned} &= \frac{1}{\sigma^2} [\sum_{i=0}^M a_i a_i r(0) + 2a_0 a_1 r(1) + 2a_0 a_2 r(2) + \dots + 2a_0 a_M r(M) + \\ &2a_1 a_2 r(1) + 2a_1 a_3 r(2) + \dots + 2a_1 a_M r(M-1) + \\ &2a_2 a_3 r(1) + 2a_2 a_4 r(2) + \dots + 2a_2 a_M r(M-2) + \\ &+ \dots + 2a_{M-1} a_M r(1)] \end{aligned}$$

$$= \frac{1}{\sigma^2} [\sum_{i=0}^M a_i a_i r(0) + 2 \sum_{n=1}^M \sum_{i=0}^{M-n} a_i a_{i+n} r(n)] =$$

$$= \frac{1}{\sigma^2} [r(0)r_a(0) + 2 \sum_{n=1}^M r_a(n)r(n)]$$

$r_a(i)$  es la correlación corta de los coeficientes  $\underline{a}$  :

$$r_a(i) = \sum_{j=0}^{M-i} a_j a_{j+i}, \quad 0 \leq i \leq M$$

Donde  $M$  es el orden del vector  $\underline{a}$ .

# Medidas de comparación de señales

Para el segundo termino de la integral:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} -(\ln S(\omega) - \ln S'(\omega))d\omega = \frac{-1}{2\pi} \int_{-\pi}^{\pi} \ln\left(\frac{S(\omega)}{S'(\omega)}\right)d\omega$$

Grenader y Szego encontraron que el resultado de esta integral es  $= \ln\left(\frac{\sigma_{\infty}^2}{\sigma_{\infty}^{\prime 2}}\right)$

donde  $\sigma_{\infty}^2$  y  $\sigma_{\infty}^{\prime 2}$  son los errores de predicción de un paso:

$$\sigma_{\infty}^2 = e^{\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln(S(\omega))d\omega}$$

$$\sigma_{\infty}^{\prime 2} = e^{\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln(S'(\omega))d\omega}$$

$$\text{Con } S'(\omega) = \frac{\sigma^2}{|A(e^{j\omega})|^2}$$

$$\frac{-1}{2\pi} \int_{-\pi}^{\pi} \ln\left(\frac{S(\omega)}{S'(\omega)}\right)d\omega = -\ln\sigma_{\infty}^2 + \ln\sigma^2$$

# Medidas de comparación de señales

Entonces la distancia de Itakura Saito:

$$\begin{aligned}d_{is}(S(\omega), \hat{S}(\omega)) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} [e^{V(\omega)} - V(\omega) - 1] d\omega \\ &= \frac{1}{\sigma^2} [r(0)r_a(0) + 2 \sum_{n=1}^M r(n)r_a(n)] - \ln\sigma_{\infty}^2 + \ln\sigma^2 - 1\end{aligned}$$

Los terminos  $\ln\sigma_{\infty}^2$  y  $\ln\sigma^2$  están relacionados con la potencia de las señales. Para el reconocimiento de palabras de voz estos terminos no son relevantes, ya que somos capaces de reconocer la voz con voz baja o voz alta, por lo tanto estos dos terminos pueden ser despreciados, lo mismo que el termino  $-1$ .

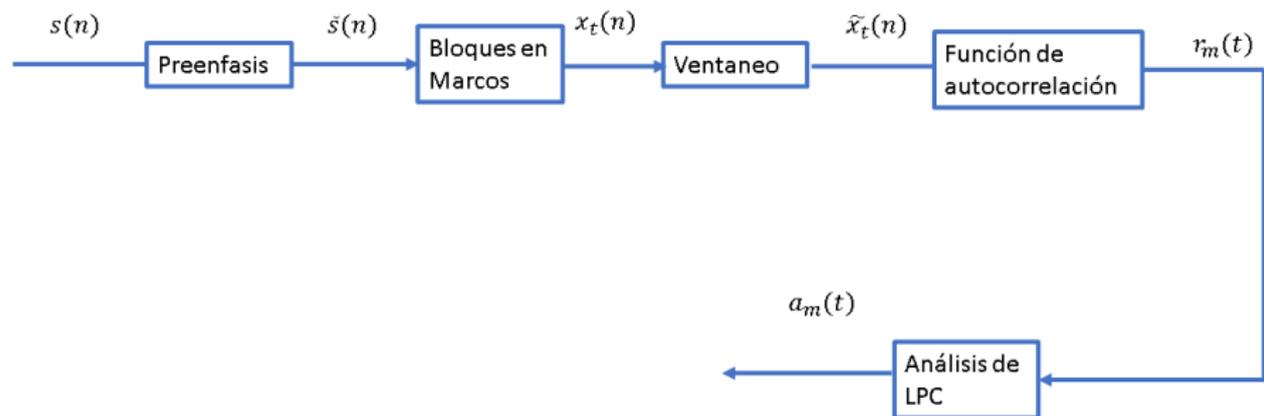
# Medidas de comparación de señales

Por lo tanto la distorsión de Itakura Saito normalizada puede ser simplificada a lo siguiente:

$$d_{is}(S(\omega), \hat{S}(\omega)) = r(0)r_a(0) + 2 \sum_{n=1}^M r(n)r_a(n)$$

Es decir se están comparando el espectro de dos señales usando solamente las correlaciones, ¡sin tener que sacar la transformada de Fourier!

# Procesamiento de Señales de Voz Utilizando Técnicas de LPC



## Filtro de Preemfasis

1.- Preenfasis.- La señal de voz  $s(n)$  es pasada a través de un filtro paso altas, para hacer que el espectro de la señal sea suavizado y así evitar problemas de precisión finita después en el procesamiento.

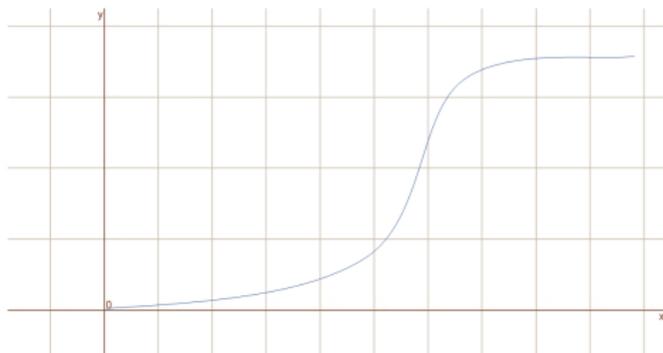
$$H(z) = 1 - \tilde{a}z^{-1}, \quad 0.9 \leq \tilde{a} \leq 1.0$$

$$H(e^{j\omega T}) = 1 - \tilde{a}e^{-j\omega T} = 1 - \tilde{a}(\cos(\omega T) - j\sin(\omega T)) \quad (1)$$

$$\begin{aligned} |H(e^{j\omega T})| &= \sqrt{(1 - \tilde{a}\cos(\omega T))^2 + \tilde{a}^2\sin^2(\omega T)} \\ &= \sqrt{1 - 2\tilde{a}\cos\omega T + \tilde{a}^2\cos^2(\omega T) + \tilde{a}^2\sin^2(\omega T)} \\ &= \sqrt{1 - 2\tilde{a}\cos(\omega T) + \tilde{a}^2} \end{aligned}$$

$$\text{Con } \tilde{a}^2 \approx 1; \quad = \sqrt{2 - 2\tilde{a}\cos(\omega T)} = \sqrt{2} \cdot \sqrt{1 - \tilde{a}\cos(\omega T)}$$

## Filtro de Preemfasis



$$\tilde{S}(z) = H(z)S(z) = (1 - \tilde{a}z^{-1})S(z) = S(z) - \tilde{a}z^{-1}S(z)$$

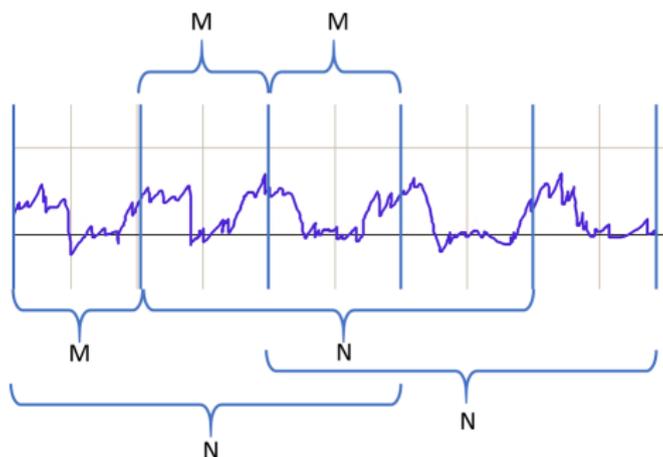
Por lo tanto:

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1)$$

## 2.- Formación de bloques

$$x_l(n) = \tilde{s}(Ml + n), \quad n = 0, 1, \dots, N-1$$

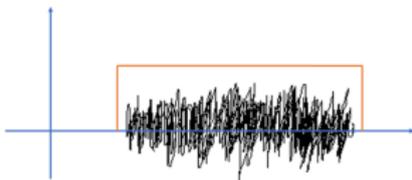
$$l = 0, 1, \dots, < L$$



El segundo bloque de N muestras está separado M muestras del primero.

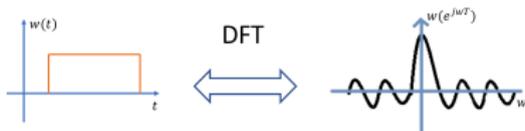
3.- Se aplica una ventana a la señal para minimizar el efecto de las discontinuidades de la señal al principio y al final de cada marco.

$$\tilde{x}_l(n) = x_l(n) \cdot w(n), \quad 0 \leq n \leq N - 1$$

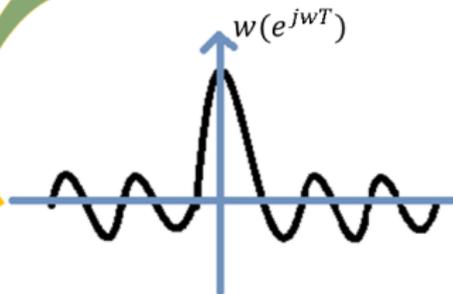


Multiplicar la señal en el dominio de la frecuencia tiene como efecto la convolución de sus transformadas discretas de Fourier:

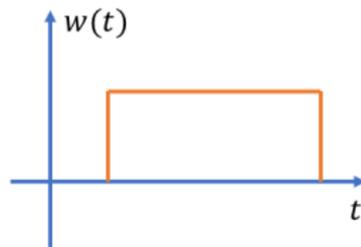
$$\tilde{X}_l(e^{j\omega t}) = X(e^{j\omega t}) * W(e^{j\omega t})$$



Se quiere una ventana que no modifique el espectro.  
Una ventana típica es la ventana de Hamming



DFT



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

#### 4.- Función de Autocorrelación

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m), \quad m = 0, 1, \dots, M$$

donde el valor de M, es el orden de los coeficientes de LPC. Valores típicos de M estan entre 8 y 16, con M=8 el valor más usado.

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

¿Cómo saber el valor de M a usar?

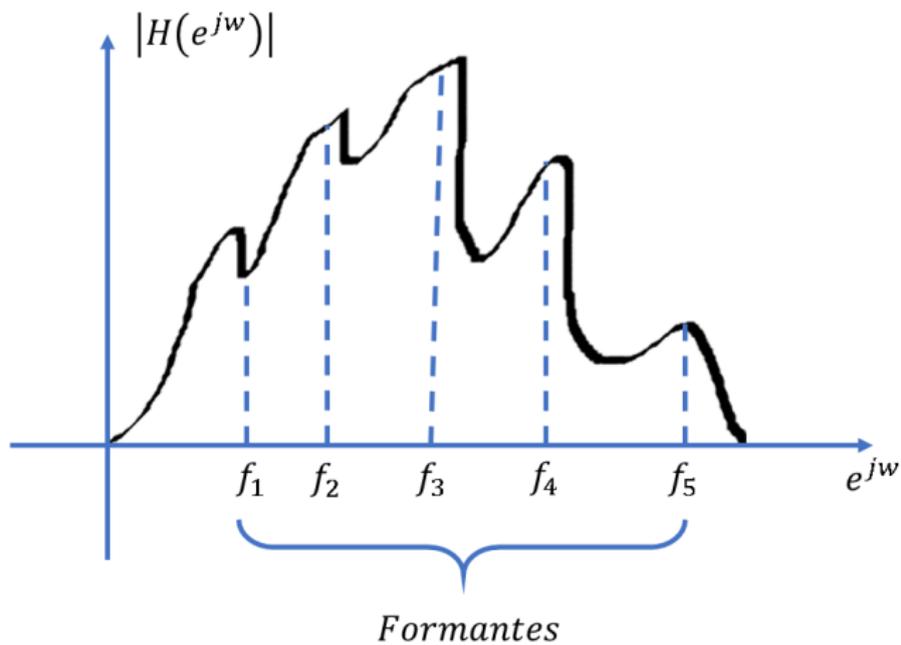
$$H(z) = \frac{\sigma}{1 + \sum_{k=1}^M a_k z^{-k}} = \frac{\sigma}{\prod_{k=0}^{M/2} (1 - z_k z^{-1})(1 - z_k^* z^{-1})}$$

Donde  $z_k$  y  $z_k^*$  son las raíces del polinomio  $\underline{a} \cdot \underline{z}^{-k}$

El espectro de  $H(z)$  tiene una indeterminación cuando  $z_k = z$ :

$$1 - z_k z^{-1} \Rightarrow z_k = z = re^{j\omega_k} \Rightarrow |H(e^{j\omega})| = \infty$$

En las frecuencias donde ocurre esto se les conoce como formantes y este es un indicador del número de coeficientes que se necesitan utilizar.



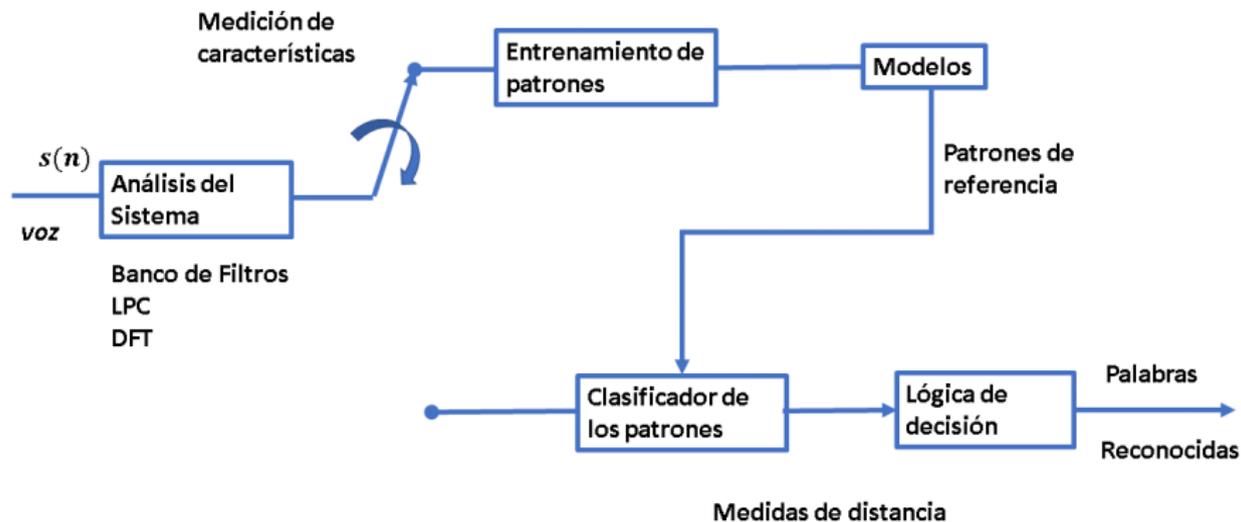


Figura: Diagrama de bloques de un sistema de reconocimiento de voz