Lección 3:
Sistemas de Reconocimiento de Voz

Dr. Jesús Savage
Dr. Carlos Rivera

26 de abril de 2021

Índice

Medidas de comparación de señales

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

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 un 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: $t - l = \tau \quad \rightarrow dt = d\tau$

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

Por lo tanto:

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

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

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:

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

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

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:

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

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

Recordar que:

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

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

Por lo tanto:

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

Medidas de comparación de señales

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

$$d_{is}(S(\omega), \hat{S}(\omega)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{S(\omega)}{S'(\omega)} d\omega = \frac{1}{2\pi\sigma^2} \int_{-\pi}^{\pi} S(\omega)|A(e^{j\omega})|^2 d\omega$$

$$= \frac{1}{2\pi\sigma^2} \int_{-\pi}^{\pi} S(\omega) \sum_{i=0}^{M} a_i e^{j\omega i} \cdot \sum_{k=0}^{M} a_k \cdot e^{-j\omega k} d\omega$$

$$= \frac{1}{2\pi\sigma^2} \sum_{i=0}^{M} a_i \cdot \sum_{k=0}^{M} a_k \int_{-\pi}^{\pi} S(\omega) \cdot e^{j\omega(i-k)} d\omega$$

$$= \frac{1}{\sigma^2} \sum_{i=0}^{M} a_i \cdot \sum_{k=0}^{M} a_k r(i-k) = \frac{1}{\sigma^2} [\underline{a}^T R \underline{a}] =$$

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

$r_a(i)$ 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:

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

Grenader y Szego encontraron que el resultado de esta integral es

$$= ln(\frac{\sigma_{\infty}^2}{\sigma_{\infty}^{2'}})$$

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

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

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

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

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

Medidas de comparación de señales

Entonces la distancia de Itakura Saito:

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

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

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.

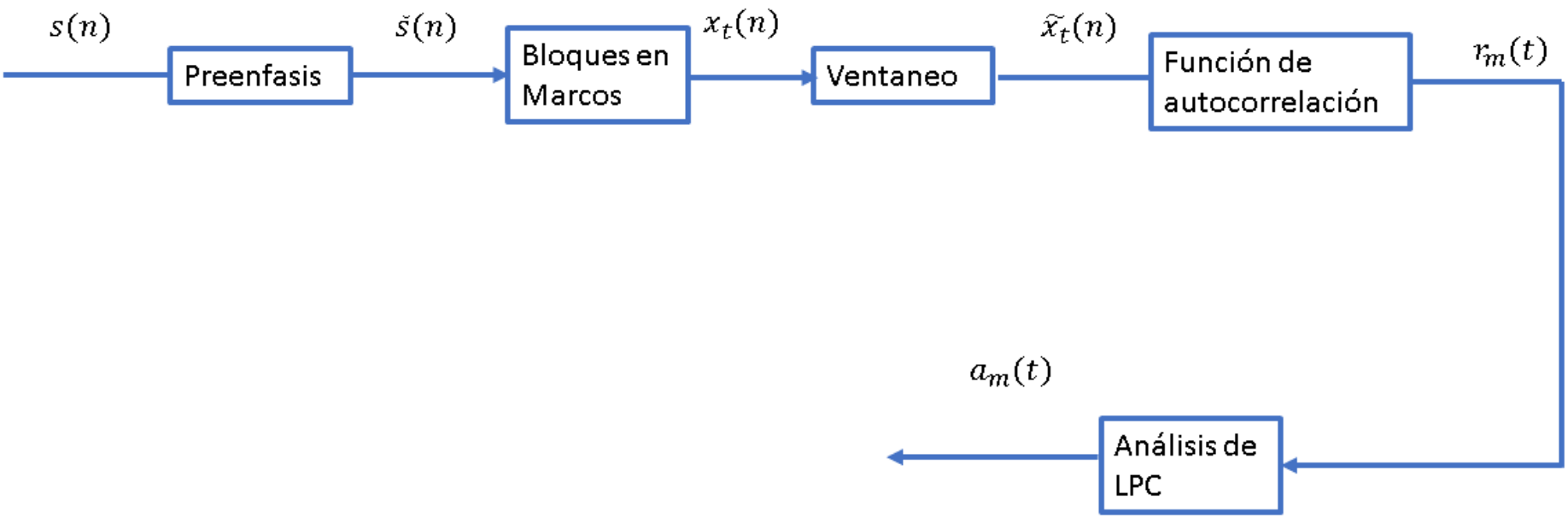
Medidas de comparación de señales

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 asi evitar problemas de precision finita despues en el procesamiento.

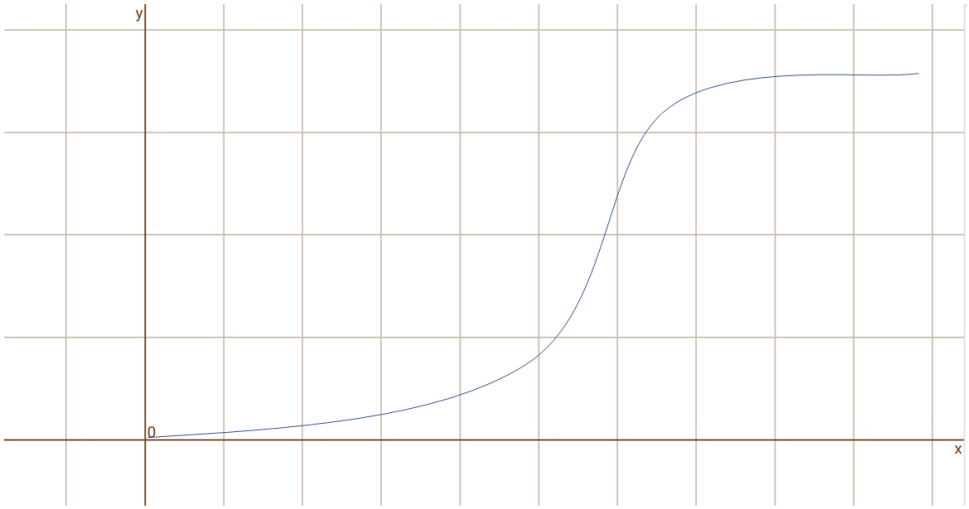
$$H(z) = 1 - \tilde{a}z^{-1}, \qquad 0.9 \leq \tilde{a} \leq 1.0$$

$$H(e^{jwT}) = 1 - \tilde{a}e^{-jwT} = 1 - \tilde{a}(cos(wT) - jsen(wT)) \tag{1}$$

$$\begin{aligned} \left|H(e^{jwT})\right| &= \sqrt{(1 - \tilde{a}cos(wT))^2 + \tilde{a}^2 sen^2(wT))} \\ &= \sqrt{1 - 2\tilde{a}coswT + \tilde{a}^2 cos^2(wT) + \tilde{a}^2 sen^2(wT))} \\ &= \sqrt{1 - 2\tilde{a}cos(wT) + \tilde{a}^2} \end{aligned}$$

Con $\tilde{a}^2 \approx 1$; $\qquad = \sqrt{2 - 2\tilde{a}cos(wT)} = \sqrt{2} \cdot \sqrt{1 - \tilde{a}cos(wT)}$

Filtro de Preemfasis



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

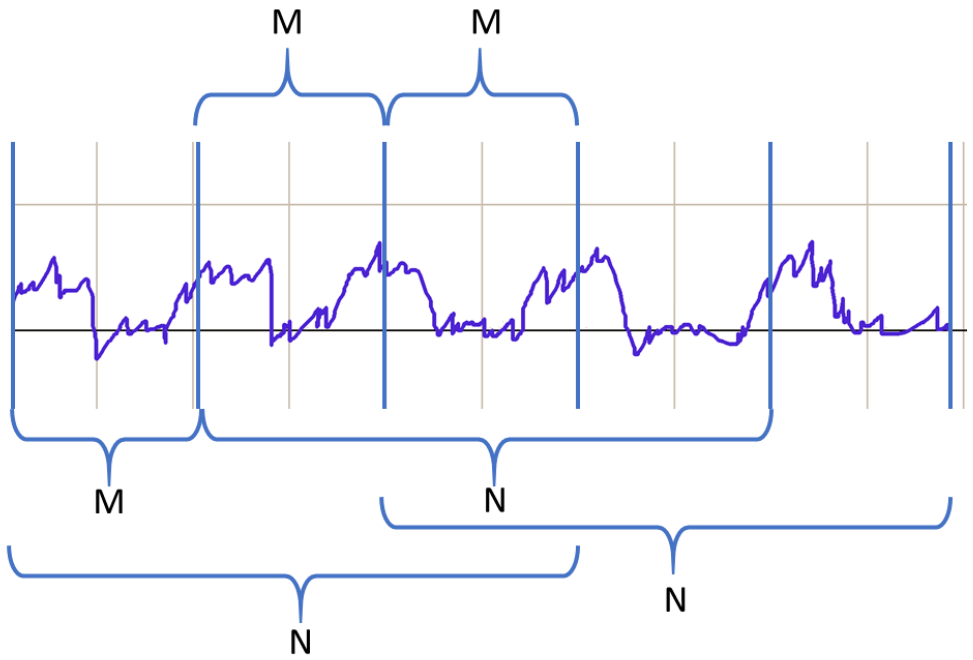
Por lo tanto:

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

2.- Formación de bloques

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

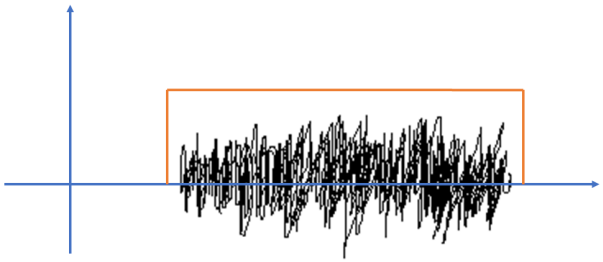
$$l = 0, 1, \cdots, < L$$



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

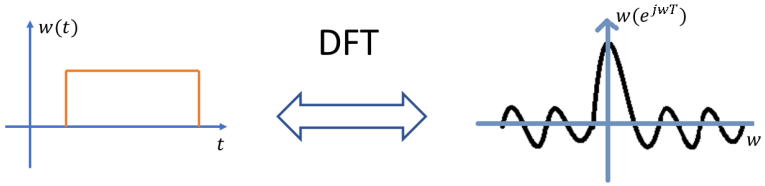
3.- Se aplica una ventana a la señal para minimizar el efecto de las discontinuidades de la señal al principio y al final de cada marco.

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



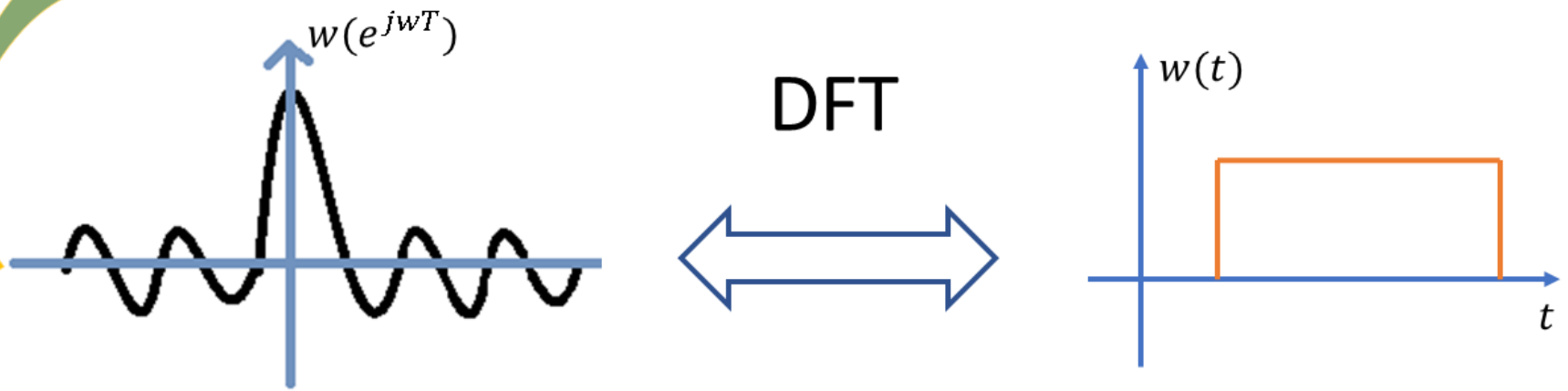
Multiplicar la señal en el dominio de la frecuencia tiene como efecto la convolución de sus transformadas discretas de Fourier:

$$\tilde{X}_l(e^{jwt}) = X(e^{jwt}) * W(e^{jwt})$$



Se quiere una ventana que no modifique el espectro.

Una ventana típica es la ventana de Hamming



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

4.- Función de Autocorrelación

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

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?

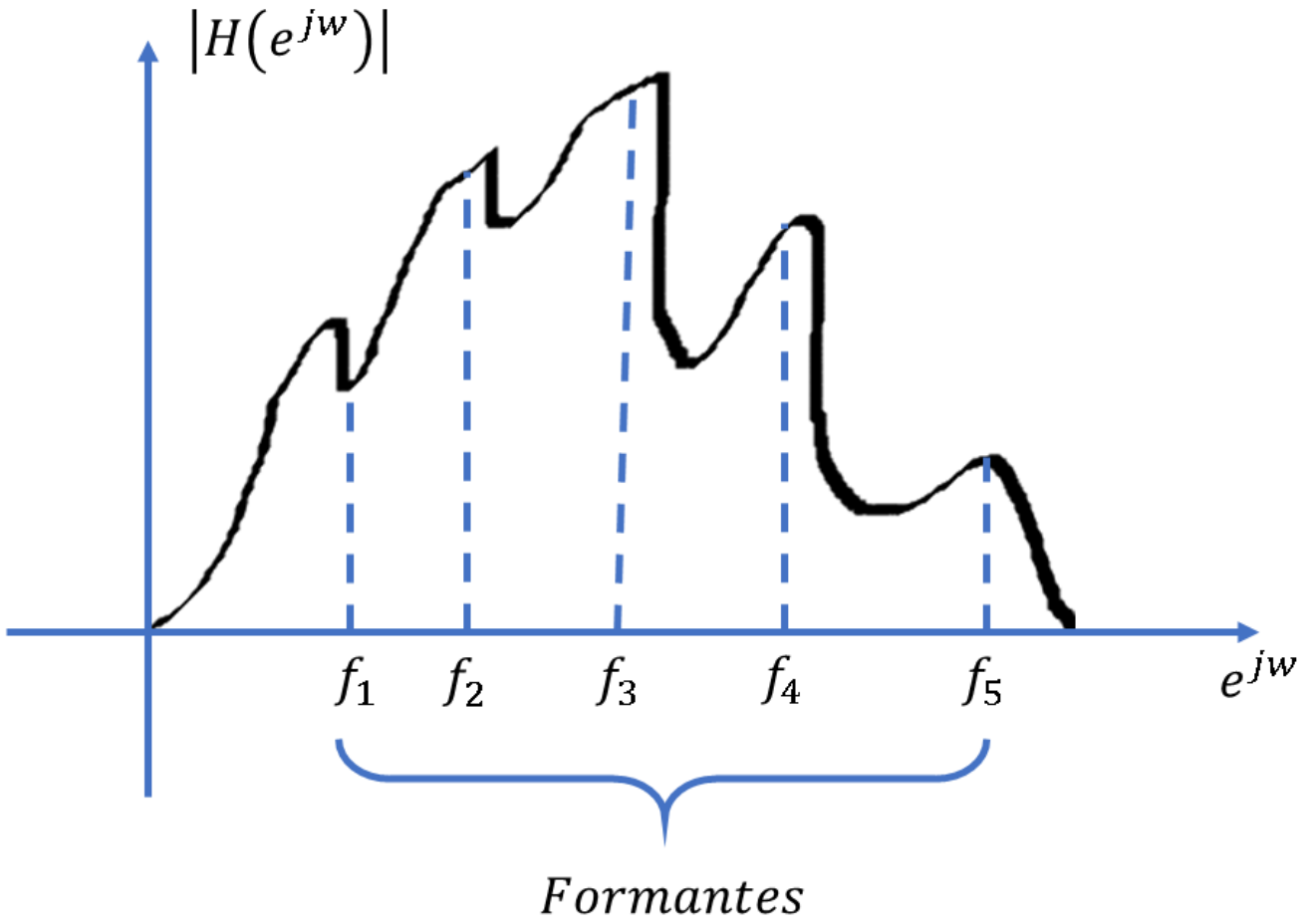
$$H(z) = \frac{\sigma}{1 + \sum_{k=1}^{M} a_k z^{-k}} = \frac{\sigma}{\prod_{k=0}^{M/2}(1 - z_k z^{-1})(1 - z_k^* z^{-1})}$$

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

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

$$1 - z_k z^{-1} \quad \Rightarrow \quad z_k = z = re^{jw_k} \Rightarrow \left|H(e^{jw})\right| = \infty$$

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


$|H(e^{jw})|$
f_1
f_2
f_3
f_4
f_5
e^{jw}
Formantes

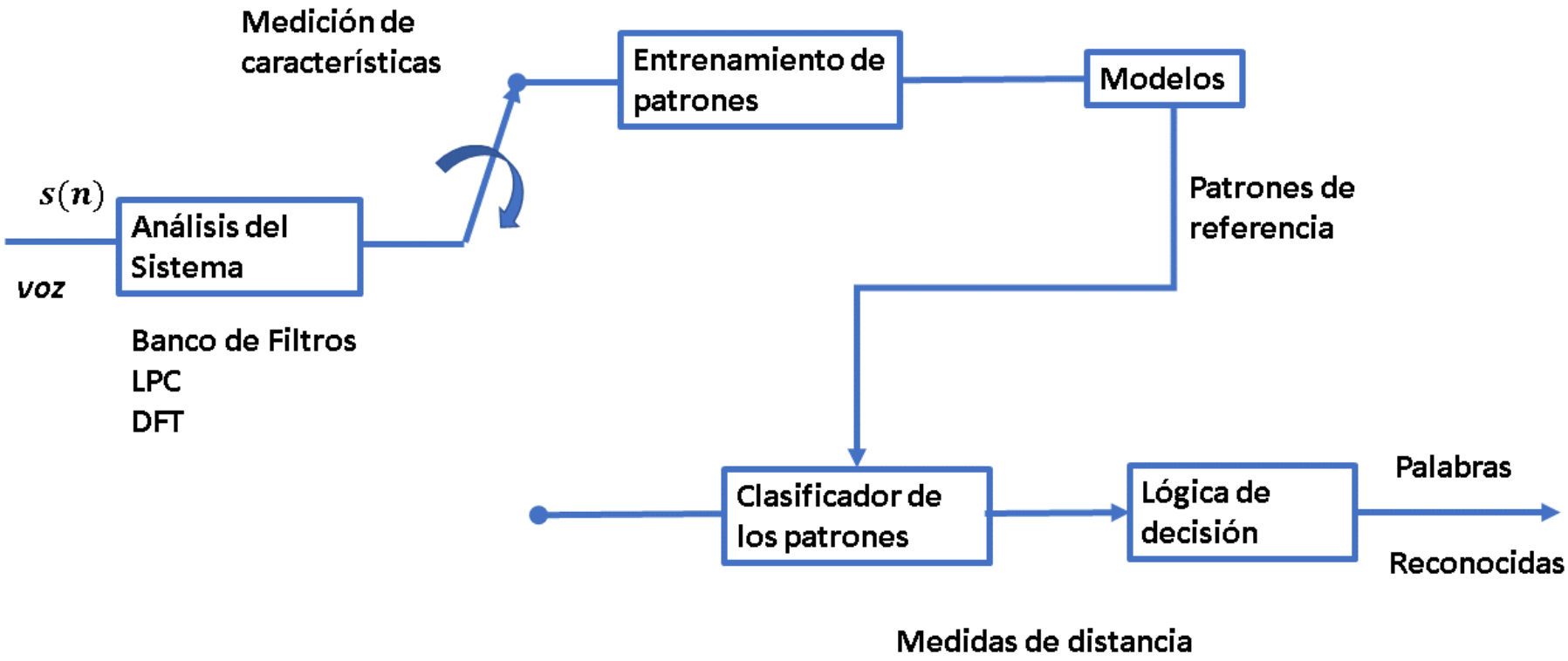


Figura: Diagrama de bloques de un sistema de reconocimiento de voz