LPC para Voz humana

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + G u(n)$$

AR - MODEL (IIR Filter)

$$x(z) = \sum_{k=1}^{p} a_k x(z) z^{-k} + G u(z)$$

$$x(z)\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right) = G u(z)$$

$$H(z) = \frac{x(z)}{u(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

ALL POLE MODEL FOR VOCAL TRACT.

Predicción LINEAL:

$$\hat{x}(n) = \sum_{k=1}^{p} a_k x(n-k)$$

predecir muestra actual de datos anteriores

$$e(n) = x(n) - \hat{x}(n)$$

$$= x(n) - \sum_{k=1}^{p} a_k x(n-k) = G u(n)$$

OBS: El error de predicción es la entrada al sistema, i.e, corresponde al sonido que se produce en las cuerdas vocales

Una estimación optima de $\hat{x}(n)$ es aquella que haga que $E[e^2(n)]$ sea mínimo (i.e, MMSE)

$$E[e^2(n)] = E[(x(n) - \hat{x}(n))^2]$$

$$= E[x^2(n) - 2x(n)\hat{x}(n) + \hat{x}^2(n)]$$

es conveniente expresar

$$\hat{x}(n) = \sum_{k=1}^{p} a_k x(n-k) = \underline{a}^T \underline{x}_n \quad \text{donde}$$

$$\underline{a} = [a_1, a_2, \cdots, a_p]^T$$

$$\underline{x}_n = [x(n-1), x(n-2), \cdots, x(n-p)]^T$$

OBS) $\hat{x}(n) = \underline{a}^T \underline{x}_n$ es 1×1

$= \underline{x}_n^T \underline{a}$ es equivalente

$$\Rightarrow E[e^2(n)] = E[x^2(n) - 2x(n)\underline{a}^T\underline{x}_n + \underline{a}^T\underline{x}_n\underline{x}_n^T\underline{a}]$$

$$= E[x^2(n)] - 2\underline{a}^T \underbrace{E[x(n)\underline{x}_n]}_{\underline{r}_x} + \underline{a}^T \underbrace{E[\underline{x}_n\underline{x}_n^T]}_{R_x}\underline{a}$$

$\underline{r}_x$: $(p \times 1)$ — R_x : $(p \times p)$

$$\underline{r}_x = \begin{bmatrix} E[x(n)\,x(n-1)] \\ E[x(n)\,x(n-2)] \\ \vdots \\ E[x(n)\cdot x(n-p)] \end{bmatrix} = \begin{bmatrix} r_{xx}(1) \\ r_{xx}(2) \\ \vdots \\ r_{xx}(p) \end{bmatrix}$$

$$\underline{R_x} = \begin{bmatrix} E[x(n-1)x(n-1)] & E[x(n-1)x(n-2)] & \cdots & E[x(n-1)x(n-p)] \\ E[x(n-2)x(n-1)] & E[x(n-2)x(n-2)] & \cdots & \\ \vdots & & & \\ E[x(n-p)x(n-1)] & \cdots & & E[x(n-p)x(n-p)] \end{bmatrix}$$

$$= \begin{bmatrix} r_{xx}(0) & r_{xx}(1) & \cdots & r_{xx}(p-1) \\ r_{xx}(1) & r_{xx}(0) & & \\ \vdots & & \ddots & \\ r_{xx}(p-1) & & & r_{xx}(0) \end{bmatrix}$$

Toeplitz y simétrica

OBS) $E[x(n)x(m)] = r_{xx}(n,m)$ autocorrelación

si x es WSS $\Rightarrow$ $r_{xx}(n,m) = r_{xx}(|n-m|)$

$$\Rightarrow E[e^2(n)] = r_{xx}(0) - 2\underline{a}^T \underline{r_x} + \underline{a}^T \underline{R_x}\,\underline{a}$$

$\nabla E[e^2(n)] = 0 \Rightarrow$ $-2\underline{r_x} + 2\underline{R_x}\,\underline{a} = 0$

(derivar con respecto a "a")

$$\Rightarrow \boxed{\underline{R_x}\,\underline{a} = \underline{r_x}}$$

Esto se resuelve eficientemente con Levison - Durbin

OBS] • los coeficientes "a" sólo representan la predicción lineal, pero no el filtro AR.

Este filtro se compone por:

$$\hat{a} = [1, -a_1, -a_2, \dots, -a_p]$$

• la estimación de r_{xx} no es trivial. Una forma es usar

$$\hat{r}_{xx}[m] = \frac{1}{N} \sum_{n=0}^{N-m-1} x[n] \cdot x[n+m] \qquad 0 \le m \le p$$

lo que se obtiene con "xcorr" en matlab.

• Una estimación adecuada de r_{xx} requiere de al menos 6 ciclos de la señal $x[n]$

¿Para qué se usa LPC en voz humana?

- Telefonía celular y VoIP : CELP y variaciones
- Reconocimiento del habla : MFCC y MFLPC
- Reconocimiento de locutor : MFCC y c[n]
- Aplicaciones clínicas : Filtrado inverso : e[n]
- Aplicaciones Fonéticas : Frecuencia de LP polos
- Síntesis Fisiológica de voz : mejora sobre sintesis con concatenación
- Reducción de ruido : Kalman LPC

OTROS CAMPOS:

- Series de tiempo y señales : modulaciones que afectan a señales que pueden ser estimadas y removidas con LPC

Dificultades típicas :

- Glotis varía su condición de borde temporalmente (no es un tubo cerrado)
- Resonancias nasales → ceros en H(z)
- Acoplamiento resonancias subglotales → ceros en H(z) y polos ↳ perturbaciones
- Acoplamiento no lineal entre Fuente y Filtro
- Problemas de estimación de LPC en condiciones de nivel de Fondo alto.
- El residual o Fuente en las cuerdas no siempre está en las cuerdas ni tiene mucho valor Fisiológicamente o Físicamente relevante.
- No siempre LPC está bien definido para Consonantes e impulsividades de la voz
- La Fuente puede ceros y polos en su estructura y es difícil diferenciar éstos de los efectos del tracto vocal.

TRANSFORMADA DISCRETA DE COSENO

$$X_{DCT}^{(N)}(K) = 2 \sum_{n=0}^{N-1} x(n) \cos\left(\frac{\pi K}{2N}(2n+1)\right) \qquad K = 0, 1, \dots N-1$$

(DCT-II)

$$= 2 \sum_{n=0}^{N-1} x(n) \cos\left(\frac{2\pi K}{2N}\left(n+\frac{1}{2}\right)\right)$$

OBS]

$$\text{re}\{X^{(2N)}(K)\} = \text{re}\left\{\sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi K n}{N}}\right\}$$

PARTE REAL DE UNA DFT

$$= \sum_{n=0}^{N-1} x(n) \cos\left(\frac{2\pi K}{N} n\right)$$

RELACIÓN ENTRE DCT Y DFT

$$\Rightarrow \quad X_{DCT}^{(N)}(K) = e^{j\frac{2\pi K}{2N}} X^{(2N)}(K) + e^{-j\frac{2\pi K}{2N}} X^{(2N)}(K)$$

OBS]

- DCT es siempre real
- Tiene menores discontinuidades (asume simetría por lo que hace que la energía se distribuya en pocos valores de K (menos que en la DFT))
- Se utiliza mucho en compresión de señales por este motivo.

¿ De qué sirve CEPSTRUM ?

Supongan de una señal esta filtrada de modo que

$$x[n] = u[n] * h[n]$$
$$X(w) = U(w) \cdot H(w)$$

$$\log(X(w)) = \log(U(w) \cdot H(w))$$
$$= \log(U(w)) + \log(H(w))$$

OBS] • Bajo ciertas condiciones las características de $U(w)$ y $H(w)$ son suficientemente distintas para ser separadas directamente por el logaritmo (i.e., no tienen compartir cruzadas en sus rangos de frecuencia)

- Si $h[n]$ es de bajo orden, su energía se concentra en los primeros coff. cepstrum
- Si $u[n]$ oscila a mayor frecuencia, la energía de $u[n]$ se contendrá en componentes superiores del cepstrum
- Esto sucede en voz y análisis de señales sísmicas.

$$\Rightarrow \quad \hat{h}[n] = \text{cepstrum real}(1 : L)$$
$$\hat{x}[n] = \text{cepstrum real}(L+1 : \tfrac{N}{2})$$

OBS] • cepstrum es simétrico con respecto a su punto medio $(\frac{N}{2})$.
• L es el largo del "lifter", N es el largo de la señal.