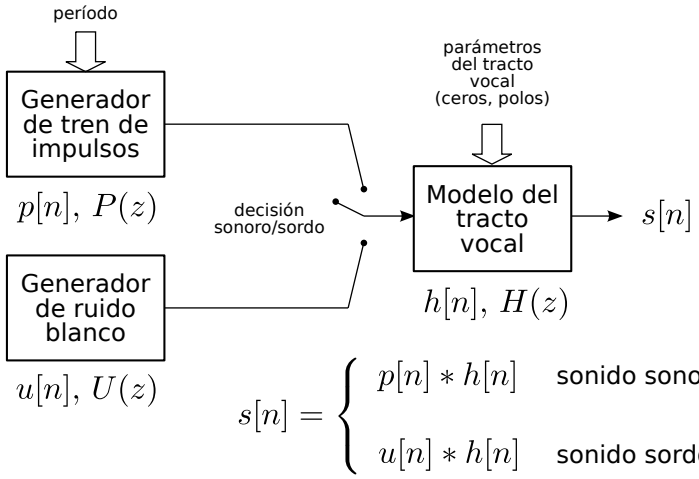
Procesamiento digital de señales de audio

Descomposición homómorfica

Instituto de Ingeniería Eléctrica
Facultad de Ingeniería

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Modelo en tiempo discreto del mecanismo de producción de la voz



Modelo general

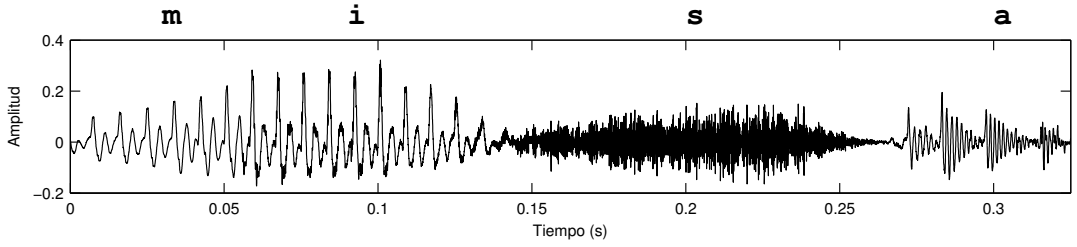
La señal de voz se representa como la salida de un sistema lineal variante en el tiempo.

Modelo en tiempo discreto del mecanismo de producción de la voz

Modelo básico

En períodos cortos de tiempo, cada fragmento de la señal de voz puede modelarse como la salida de un LTI alimentado con

- un tren de pulsos cuasi-periódico (sonidos sonoros)
- ruido blanco (sonidos sordos)



En la práctica los parámetros del modelo no se conocen, solo se conocen las muestras de la señal de voz.

Deconvolución

Objetivo

- El objetivo es inferir los parámetros del modelo dada la señal de voz.

$$\text{señal de voz} = \text{excitación} * \text{respuesta del sistema}$$

- Hay que deconvolucionar la señal de voz en la excitación y la respuesta del sistema.
- *Deconvolución ciega* (blind deconvolution):
  no se conoce ninguna de las dos señales que integran la convolución, solo se conoce el resultado de la convolución.
- Técnicas clásicas:
  - Codificación por predicción lineal (LPC, Linear Predictive Coding)
    - Paramétrica: se asume un modelo de la respuesta del sistema
  - Filtrado homomórfico
    - No paramétrica

Deconvolución

Análisis de voz

- Estimar los parámetros de la señal de voz y su evolución en el tiempo:
  - frecuencia fundamental: frecuencia de la excitación en el modelo.
  - frecuencia de las formantes: frecuencias de resonancias del tracto vocal.
- Los parámetros aparecen explícitamente en la señal de excitación y en la respuesta del tracto vocal.

Aplicaciones

- Codificación de voz a baja tasa de bits (Homomorphic vocoder).
- Reconocimiento del habla a través de la estimación de formantes.
- Modificación de audio (pitch shifting, time stretching).
- Clasificación de señales de audio.

Filtrado homomórfico [Oppenheim, 1965]

Combinación de señales

- Señales combinadas aditivamente: pueden ser separadas si su soporte temporal o espectral es disjunto.
- Señales combinadas no linealmente:
  - por convolución: excitación y respuesta del sistema en señal de voz
  - por multiplicación: atenuación variable en el tiempo al transmitir por un canal

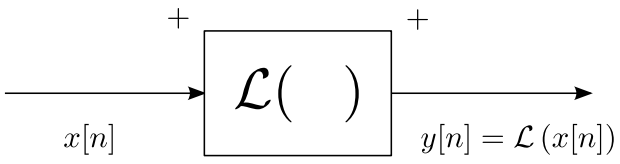
Estrategia

Fitrado homomórfico de señales combinadas por convolución consiste en:

1. Transformar las señales combinadas por convolución a señales combinadas aditivamente.
2. Aplicar un filtro lineal para la separación.
3. Aplicar la transformación inversa para volver al dominio original.

Principio de superposición

Un *sistema lineal* cumple el principio de superposición



$x_1[n] + x_2[n]$ $\qquad$ $\mathcal{L}(x_1[n]) + \mathcal{L}(x_2[n])$

$\alpha x[n]$ $\qquad$ $\alpha\mathcal{L}(x[n])$

Observación

- La salida de una combinación aditiva de señales elementales resulta en la combinación aditiva de las salidas individuales.
- Esto se indica con el signo de $+$ en la entrada y en la salida en la figura: la suma de entradas se transforma en suma de salidas.

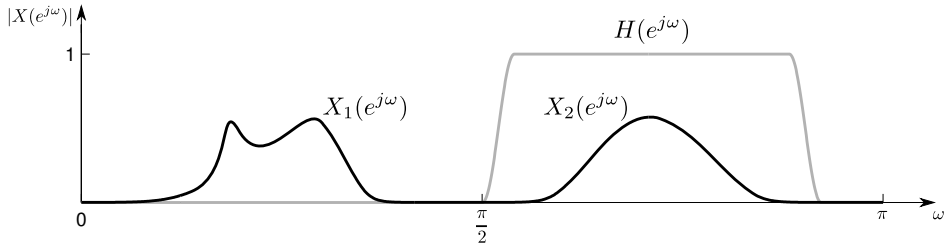
Principio de superposición

Una consecuencia del principio de superposición es la capacidad de los sistema lineales de separar señales con espectros disjuntos.

Ejemplo

- $x_1[n]$ y $x_2[n]$ con espectros disjuntos, $h[n]$ filtro pasaltos
- Si la entrada al filtro es $x_1[n] + x_2[n]$, la salida es

$$\begin{aligned} y[n] &= h[n] * (x_1[n] + x_2[n]) \\ &= h[n] * x_2 \\ &= x_2[n] \end{aligned}$$



Principio de superposición generalizado

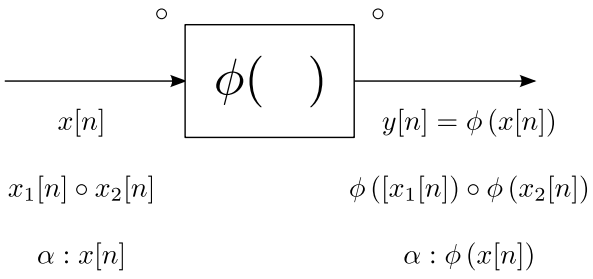
Definición

Sean los operadores "$\circ$" y ":", donde

- $\circ$ - regla de combinación de señales, $x[n] = x_1[n] \circ x_2[n]$
- $:$ - operador de multiplicación generalizado,

el sistema ϕ cumple el principio de superposición generalizado si

$$\begin{aligned} \phi(x_1[n] \circ x_2[n]) &= \phi(x_1[n]) \circ \phi(x_2[n]) \\ \phi(\alpha : x[n]) &= \alpha : \phi(x[n]) \end{aligned}$$



Sistema homomórfico

Definición

Un sistema se dice homomórfico para la regla de combinación $\circ$ y de producto generalizado : si cumple el principio de superposición con esas operaciones.

Observaciones

- Un sistema lineal es homomórfico para la regla de combinación $+$.
- Para la deconvolución de la señal de voz interesa estudiar los sistemas homomórficos para convolución.
  - La operación de combinación es la convolución: $\circ = *$.
  - La operación de multiplicación generalizada no se empleará.

Sistema homomórfico para convolución

Definición

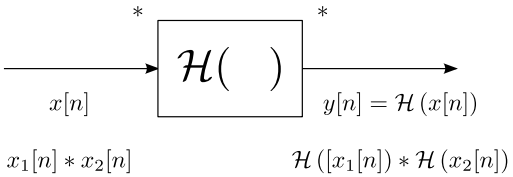
Sea el sistema $\mathcal{H}$ que produce la salida

- $y_1[n]$ cuando la entrada es $x_1[n]$, con $y_1[n] = \mathcal{H}(x_1[n])$
- $y_2[n]$ cuando la entrada es $x_2[n]$, con $y_2[n] = \mathcal{H}(x_2[n])$.

El sistema $\mathcal{H}$ es homomórfico para convolución si se cumple que

$$\begin{aligned}\mathcal{H}(x_1[n] * x_2[n]) &= \mathcal{H}(x_1[n]) * \mathcal{H}(x_2[n]) \\ &= y_1[n] * y_2[n],\end{aligned}$$

es decir, si la entrada es $x_1[n] * x_2[n]$, la salida es $y_1[n] * y_2[n]$.



Filtro homomórfico para convolución

Definición

Un filtro homomórfico es un sistema homomórfico con la propiedad de que uno de los componentes (el deseado) pasa esencialmente inalterado mientras que el otro es eliminado.

Ejemplo

- sea la señal $x[n] = x_1[n] * x_2[n]$
- se quiere recuperar la señal $x_2[n]$

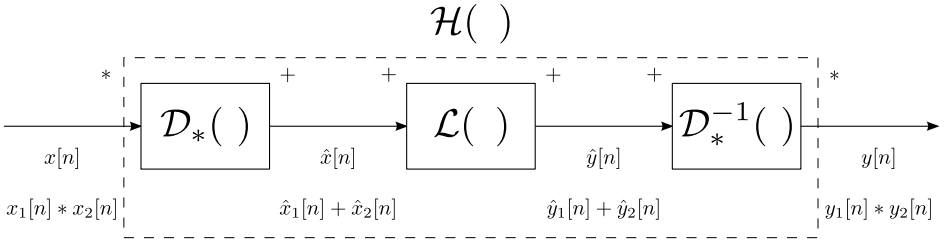
Se construye un filtro homomórfico tal que

- $\mathcal{H}(x_1[n]) \approx \delta[n]$
- $\mathcal{H}(x_2[n]) \approx x_2[n]$

De esta forma, la salida del filtro cuando la entrada es $x[n]$ es

$$y[n] = \mathcal{H}(x[n]) = \mathcal{H}(x_1[n]) * \mathcal{H}(x_2[n]) \approx \delta[n] * x_2[n] = x_2[n].$$

Representación canónica de sistemas homomórficos

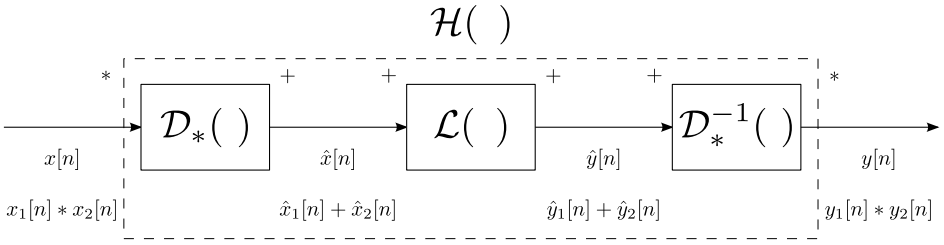


Propiedad

Todo sistema homomórfico puede representarse como la cascada de tres sistemas homomórficos.

1. el primer sistema recibe señales combinadas por convolución y devuelve señales combinadas por adición
2. el segundo sistema es un sistema lineal convencional que responde al principio de superposición
3. el tercer sistema es el inverso del primero, recibe señales combinadas por adición y devuelve señales combinadas por convolución

Representación canónica de sistemas homomórficos



Sistema característico

- $\mathcal{D}_*(\ )$: Sistema característico para deconvolución homomórfica

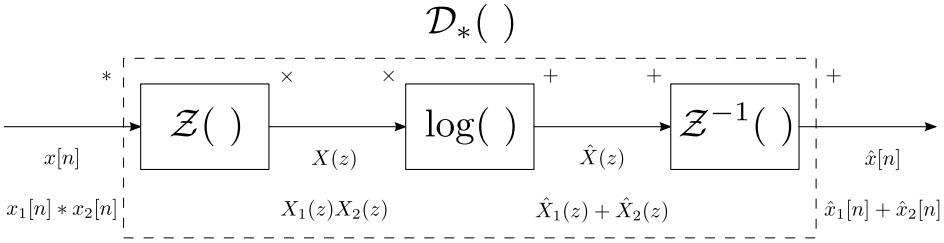
$$\mathcal{D}_*(x_1[n] * x_2[n]) = \mathcal{D}_*(x_1[n]) + \mathcal{D}_*(x_2[n]) = \hat{x}_1[n] + \hat{x}_2[n]$$

- $\mathcal{D}_*^{-1}(\ )$: Sistema característico inverso para deconvolución homomórfica

$$\mathcal{D}_*^{-1}(\hat{y}_1[n] + \hat{y}_2[n]) = \mathcal{D}_*^{-1}(\hat{y}_1[n]) * \mathcal{D}_*^{-1}(\hat{y}_2[n]) = y_1[n] * y_2[n]$$

- Ambos sistemas son fijos en la representación canónica.
- Consecuencia: el diseño de un sistema homomórfico se reduce al diseño de un sistema lineal.

Sistema característico



Definición

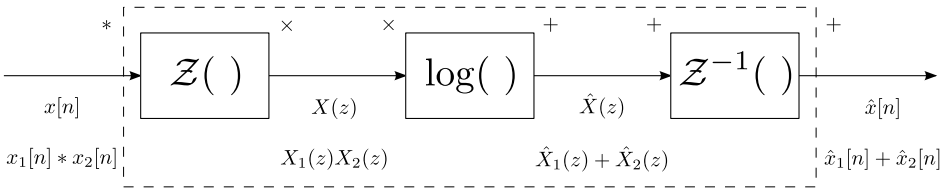
$$\mathcal{D}_*(x[n]) = \mathcal{Z}^{-1}\left(\log\left(\mathcal{Z}(x[n])\right)\right)$$

Observaciones

- La transformada $\mathcal{Z}$ mapea la convolución de secuencias en un producto.
- El logaritmo mapea el producto en una suma.
- La transformada $\mathcal{Z}$ inversa se incluye para obtener la salida en el dominio del tiempo.

Sistema característico

$\mathcal{D}_*(\ )$



Deconvolución

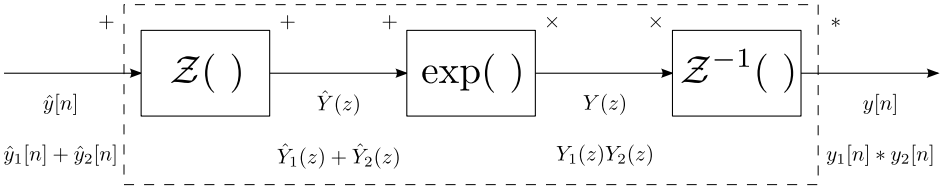
Sea la entrada $x[n] = x_1[n] * x_2[n]$. Como $X(z) = X_1(z)X_2(z)$, la salida es

$$\begin{aligned}\hat{x}[n] &= \mathcal{Z}^{-1}(\log(X(z))) \\ &= \mathcal{Z}^{-1}(\log(X_1(z)X_2(z))) \\ &= \mathcal{Z}^{-1}(\log(X_1(z)) + \log(X_2(z))) \\ &= \mathcal{Z}^{-1}(\log(X_1(z))) + \mathcal{Z}^{-1}(\log(X_2(z))) \\ &= \hat{x}_1[n] + \hat{x}_2[n],\end{aligned}$$

con $\hat{x}_1[n]$ y $\hat{x}_2[n]$ las salidas correspondientes a las entradas $x_1[n]$ y $x_2[n]$.

Sistema característico inverso

$\mathcal{D}_*^{-1}(\ )$



Definición

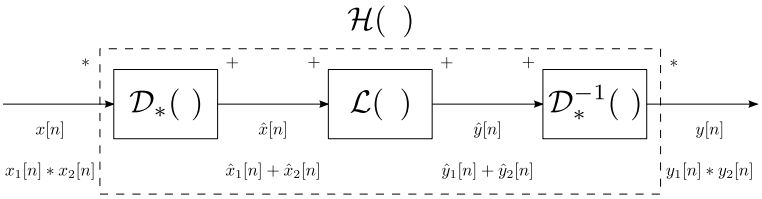
$$\mathcal{D}_*(\hat{y}[n]) = \mathcal{Z}^{-1}\left(\exp\left(\mathcal{Z}(\hat{y}[n])\right)\right)$$

Observaciones

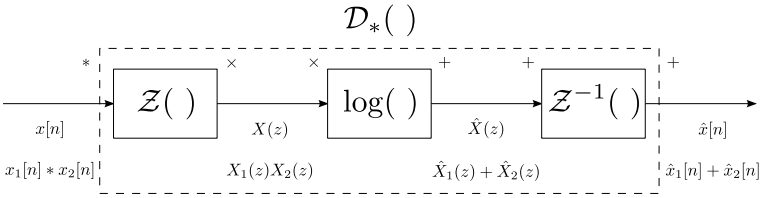
- Si $\hat{x}[n] = \mathcal{D}_*(x[n])$, entonces $x[n] = \mathcal{D}_*^{-1}(\hat{x}[n])$.

Sistema homomórfico para convolución

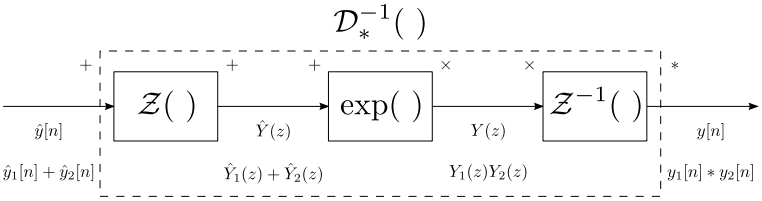
Representación canónica de sistema homomórfico para convolución



Sistema característico



Sistema característico inverso



Cepstrum complejo

Definición

- La salida del sistema característico se denomina cepstrum complejo.
- El cepstrum complejo puede calcularse empleando la transformada de Fourier de Tiempo Discreto,

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log(X(e^{j\omega}))e^{j\omega n} d\omega.$$

El cepstrum complejo de una secuencia es la transformada inversa de Fourier del logaritmo de la transformada de Fourier de la secuencia.

Observaciones

- El cálculo analítico del cepstrum complejo es en general mas sencillo empleando directamente la transformada $\mathcal{Z}$.
- El cálculo del cepstrum complejo involucra el logaritmo de una función compleja, $X(z)$ o $X(e^{j\omega})$.

Cepstrum complejo

Logaritmo complejo

$$\begin{aligned}\log X(e^{j\omega}) &= \log\left(|X(e^{j\omega})|e^{j\angle X(e^{j\omega})}\right)\\ &= \log\left(|X(e^{j\omega})|\right) + j\angle X(e^{j\omega})\end{aligned}$$

Observación

Si $x[n]$ es real, $X(e^{j\omega})$ es una función hermítica,

- $|X(e^{j\omega})|$ es par $\Rightarrow \mathrm{Re}\{\log X(e^{j\omega})\} = \log\left(|X(e^{j\omega})|\right)$ es par.
- $\angle X(e^{j\omega})$ es impar $\Rightarrow \mathrm{Im}\{\log X(e^{j\omega})\} = \angle X(e^{j\omega})$ es impar.

Por lo tanto $\log X(e^{j\omega})$ también es hermítica por lo que corresponde a la transformada de Fourier de una señal real.

El cepstrum complejo de una secuencia real es una secuencia real.

Consideraciones sobre el logaritmo complejo

Logaritmo complejo

- Para que el sistema característico transforme la convolución en suma, se tiene que cumplir que si $X(e^{j\omega}) = X_1(e^{j\omega})X_2(e^{j\omega})$,

$$\log X(e^{j\omega}) = \log X_1(e^{j\omega})X_2(e^{j\omega}) = \log X_1(e^{j\omega}) + \log X_2(e^{j\omega})$$

- Equivalentemente, se tiene que cumplir que

$$\begin{aligned}\log|X_1(e^{j\omega})X_2(e^{j\omega})| &= \log|X_1(e^{j\omega})| + \log|X_2(e^{j\omega})|\\ \angle X_1(e^{j\omega})X_2(e^{j\omega}) &= \angle X_1(e^{j\omega}) + \angle X_2(e^{j\omega})\end{aligned}$$

Consideraciones sobre el logaritmo complejo

Desdoblamiento de fase

Problema: la igualdad en la ecuación de la fase no necesariamente se cumple dada la ambigüedad en el argumento de un número complejo,

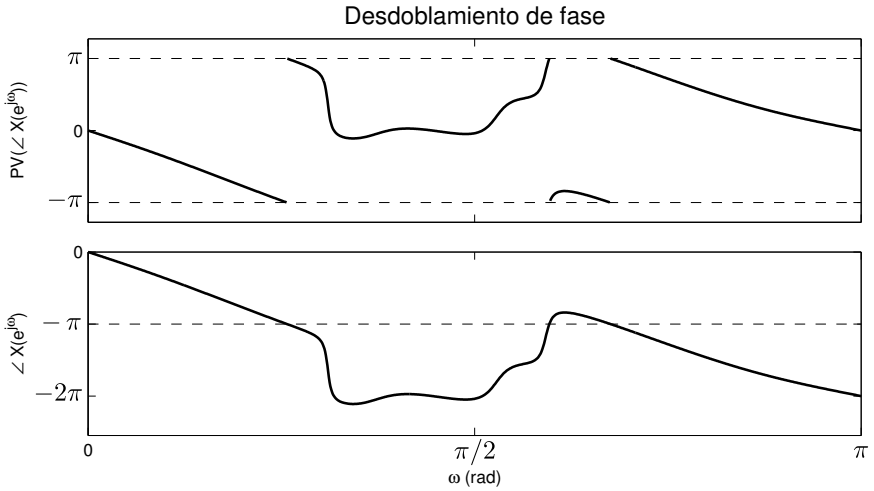
$$\angle X(e^{j\omega}) = \mathrm{PV}\left(\angle X(e^{j\omega})\right) + 2k\pi, \quad \forall k \in \mathbb{N},$$

con $\mathrm{PV}\left(\angle X(e^{j\omega})\right) \in [-\pi, \pi]$.

Solución: forzar continuidad en la fase de $X(e^{j\omega})$ (desdoblamiento de fase),

$$\text{elegir } k(\omega) \in \mathbb{N} \text{ tal que } \angle X(e^{j\omega}) = \mathrm{PV}\left(\angle X(e^{j\omega})\right) + 2k(\omega)\pi \text{ sea contínua.}$$

Consideraciones sobre el logaritmo complejo



Desdoblamiento de fase

- se elimina ambigüedad de la fase
- se cumple que $\angle X_1(e^{j\omega})X_2(e^{j\omega}) = \angle X_1(e^{j\omega}) + \angle X_2(e^{j\omega})$

Consideraciones sobre el logaritmo complejo

Eliminación de componente lineal

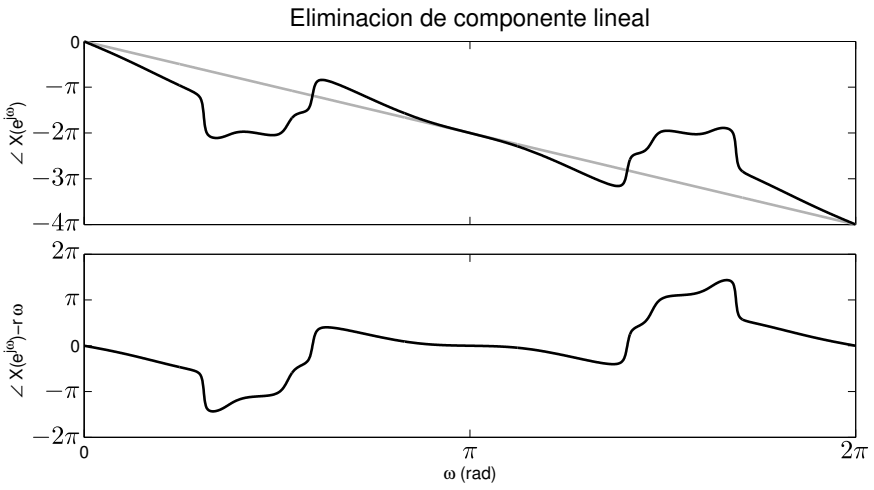
Problema: Si $x[n]$ es real,

- $\log X(e^{j\omega})$ debe ser una función hermítica:
  - $\log|X(e^{j\omega})|$ par: lo es por ser $|X(e^{j\omega})|$ par
  - $\angle X(e^{j\omega})$ impar: puede no serlo por el desdoblamiento de fase
- $\log|X(e^{j0})|$ y $\log|X(e^{j\pi})|$ reales
  - $\angle X(e^{j0}) = \angle X(e^{j\pi}) = 0$: puede no serlo por el desdoblamiento de fase

Solución: eliminar un componente lineal de la fase,

$$\angle X(e^{j\omega}) - r\omega, \quad \text{con } r = \frac{\angle X(e^{j\pi})}{\pi} \in \mathbb{N}$$

Consideraciones sobre el logaritmo complejo



Eliminación de componente lineal

- $x[n-r] \overset{\mathcal{F}}{\longleftrightarrow} X(e^{j\omega})e^{-j\omega r}$
- Se está calculando el cepstrum de $x[n-r]$ y no de $x[n]$.

Cepstrum

Cepstrum o Cepstrum real

- El cepstrum real se define como

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(|X(e^{j\omega})|)e^{j\omega n}d\omega.$$

El cepstrum real es la transformada inversa de Fourier del logaritmo de la magnitud del espectro.

- Se llama real porque se considera solo la parte real del logaritmo complejo,

$$\mathcal{F}(c[n]) = \log|X(e^{j\omega})| = \mathrm{Re}\{\log X(e^{j\omega})\} = \mathrm{Re}\{\mathcal{F}(\hat{x}[n])\}$$

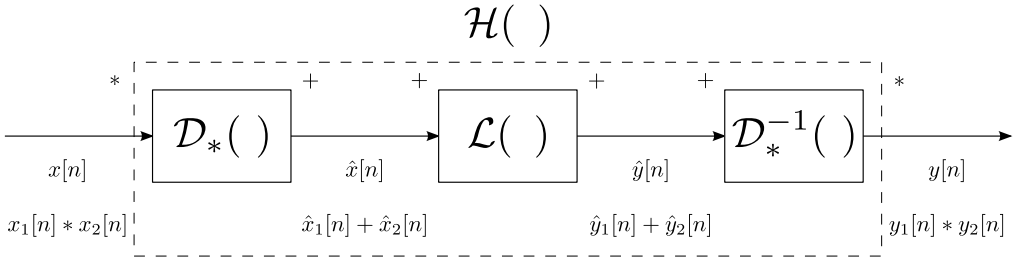
- Como la transformada de Fourier de $c[n]$ es la parte real de la transformada de Fourier de $\hat{x}[n]$, $c[n]$ es la parte par de $\hat{x}[n]$,

$$c[n] = \frac{\hat{x}[n] + \hat{x}[-n]}{2}$$

Observaciones

Dominio del cepstrum

- La aplicación de la transformada inversa de Fourier al espectro logarítmico hace que el cepstrum y el cepstrum complejo sean funciones del índice temporal n.
- El dominio temporal del cepstrum se denomina quefrency.



Diseño del filtro lineal del sistema homomórfico

- Los sistemas homomórficos para convolución solo difieren en el sistema lineal.
- La salida del sistema característico es el cepstrum complejo.
- Es necesario conocer las características del cepstrum complejo para diseñar el filtro lineal.

Análisis del cepstrum complejo

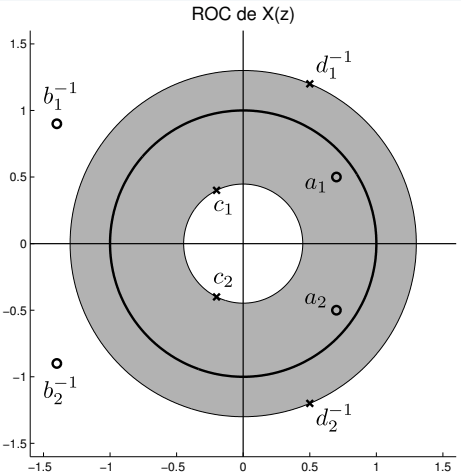
- Objetivo:
  Aplicar el cesptrum complejo en la deconvolución de señales de voz para separar la excitación de la respuesta al impulso del tracto vocal.
- Se estudiarán las características de las secuencias involucradas en el modelo del mecanismo de producción de la voz:
  - secuencias con transformada $\mathcal{Z}$ racional
  - tren de pulsos períodico

Secuencias con transformada $\mathcal{Z}$ racional

X(z) racional genérica

$$X(z) = Az^{-r}\frac{\prod_{k=1}^{M_i}(1-a_k z^{-1})\prod_{k=1}^{M_o}(1-b_k z)}{\prod_{k=1}^{N_i}(1-c_k z^{-1})\prod_{k=1}^{N_o}(1-d_k z)}$$

- $|a_k|, |b_k|, |c_k|, |d_k| < 1$
  - Ceros en a_k dentro del círculo unidad y en b_k^{-1} fuera del circulo unidad
  - Polos en c_k dentro del círculo unidad y en d_k^{-1} fuera del circulo unidad
- z^{-r} es un retardo respecto al origen temporal. Se ignora.
- $A > 0$ es un factor de ganancia.



Secuencias con transformada $\mathcal{Z}$ racional

Logaritmo complejo de X(z), $\hat{X}(z) = \log X(z)$

$$\hat{X}(z) = \log A + \sum_{k=1}^{M_i} \log(1 - a_k z^{-1}) + \sum_{k=1}^{M_o} \log(1 - b_k z) - \sum_{k=1}^{N_i} \log(1 - c_k z^{-1}) - \sum_{k=1}^{N_o} \log(1 - d_k z)$$

- Se busca que la transformada $\mathcal{Z}$ inversa $\hat{x}[n]$ sea estable.
- La región de convergencia de $\hat{X}(z)$ debe incluir el círculo unidad.
- Los términos de $\hat{X}(z)$ son de la forma
  - $\log(1 - \alpha z^{-1})$, con $|\alpha| < 1$
  - $\log(1 - \beta z)$, con $|\beta| < 1$

  y deben representar transformadas $\mathcal{Z}$ de secuencias con ROC que incluyan el círculo unidad.

Secuencias con transformada $\mathcal{Z}$ racional

Expansión en series de potencia

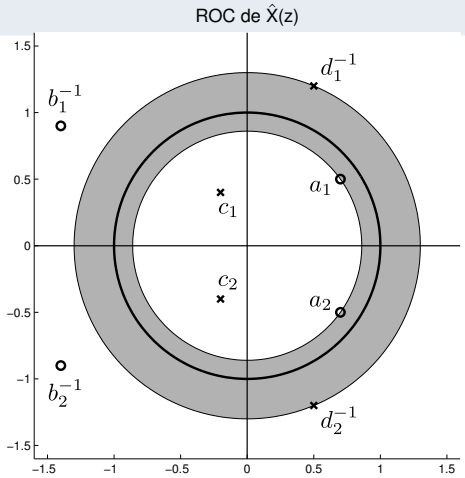
$$\log(1 - \alpha z^{-1}) = -\sum_{n=1}^{\infty} \frac{\alpha^n}{n} z^{-n}, \quad |\alpha z^{-1}| < 1 \tag{1}$$

$$\log(1 - \beta z) = -\sum_{n=1}^{\infty} \frac{\beta^n}{n} z^{n}, \quad |\beta z| < 1 \tag{2}$$

(1) $\log(1 - \alpha z^{-1})$
- Converge en $|z| > |\alpha|$
- Corresponde a una secuencia hacia adelante

(2) $\log(1 - \beta z)$
- Converge en $|z| < |\beta|^{-1}$
- Corresponde a una secuencia hacia atrás

La ROC de $\hat{X}(z)$ es el anillo determinado por los polos o ceros mas cercanos al círculo unidad.



Secuencias con transformada $\mathcal{Z}$ racional

Cepstrum complejo

$$\hat{x}[n] = \log A\delta[n] - \left[\sum_{k=1}^{M_i} \frac{a_k^n}{n} - \sum_{k=1}^{N_i} \frac{c_k^n}{n}\right] u[n-1] + \left[\sum_{k=1}^{M_o} \frac{b_k^{-n}}{n} - \sum_{k=1}^{N_o} \frac{d_k^{-n}}{n}\right] u[-n+1]$$

Observaciones

- El cepstrum complejo es la suma de exponenciales decrecientes.
- Ceros y polos dentro del círculo unidad contribuyen a la parte $n > 0$.
- Ceros y polos fuera del círculo unidad contribuyen a la parte $n < 0$.
- Es de soporte infinito aunque $x[n]$ sea causal, estable e incluso de soporte finito.
- Decrece con $|n|$ al menos como $1/|n|$.

Secuencias con transformada $\mathcal{Z}$ racional

Cepstrum complejo para secuencias de fase mínima

- $x[n]$ de fase mínima: todos los polos y ceros dentro del círculo unidad.
- $\Rightarrow \hat{x}[n]$ secuencia hacia adelante.
- Recordando que el cepstrum es la parte par del cepstrum complejo, el cepstrum complejo puede obtenerse a partir del cepstrum como

$$\hat{x}[n] = \begin{cases} 0 & n < 0 \\ c[n] & n = 0 \\ 2c[n] & n > 0 \end{cases}$$

El cepstrum complejo puede obtenerse a partir de $\log|X(z)|$ (o de $\angle X(z)$) para secuencias de fase mínima.

- Lo análogo se cumple para secuencias de fase máxima.

Tren de pulsos periódico

Tren de pulsos periódico de período N con pesos variables

$$p[n] = \sum_{r=0}^{Q} \alpha_r \delta[n - rN]$$

- La transformada $\mathcal{Z}$ es un polinomio de grado Q en z^N:

$$P(z) = \sum_{r=0}^{Q} \alpha_r z^{-rN} = \sum_{r=0}^{Q} \alpha_r \left(z^N\right)^{-r} = \prod_{r=0}^{Q-1} \left[1 - a_r \left(z^N\right)^{-1}\right]$$

- $P(z)$ consiste en el producto de términos de la forma $\left(1 - a_r \mu^{-1}\right)$ con $\mu = z^N$.

Tren de pulsos periódico

Cepstrum complejo

- Si $|a_r \mu^{-1}| < 1$ ($p[n]$ de fase mínima),

$$\begin{aligned}
\hat{P}(z) &= \log \prod_{r=0}^{Q-1} \left[1 - a_r \left(z^N\right)^{-1}\right] \\
&= \sum_{r=0}^{Q-1} \log \left[1 - a_r \left(z^N\right)^{-1}\right] \\
&= \sum_{r=0}^{Q-1} \left[-\sum_{k=1}^{\infty} \frac{a_r^k}{k} \left(z^N\right)^{-k}\right]
\end{aligned}$$

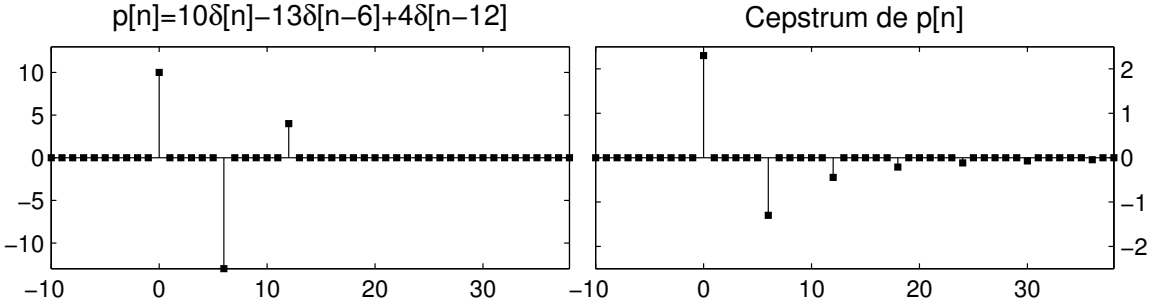
- El cepstrum complejo es

$$\hat{p}[n] = \sum_{r=0}^{Q-1} \left[-\sum_{k=1}^{\infty} \frac{a_r^k}{k} \delta[n - kN]\right]$$

Tren de pulsos periódico

Características del cepstrum complejo

- Es otro tren de pulsos periódico con el mismo período N.
- Es una secuencia hacia adelante de largo infinito.
- En el caso de tren de pulsos que no son de fase mínima el cepstrum complejo es
  - un tren de pulsos periódico con el mismo período.
  - hacia los dos lados.
  - de largo infinito.



Ejemplos

Secuencia con transferencia racional

$$H(z) = \frac{(1 - bz)(1 - b^* z)}{(1 - cz^{-1})(1 - c^* z^{-1})}, \quad \text{con } |b|, |c| < 1$$

- Par de ceros conjugados fuera del círculo unidad.
- Par de polos conjugados dentro del círculo unidad.

Cepstrum complejo

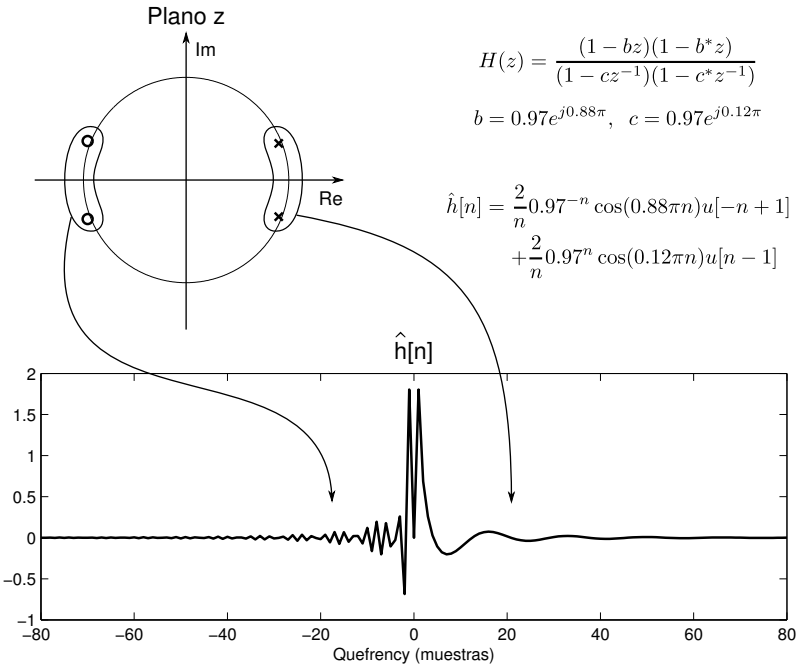
Si

$$b = |b| e^{j\theta_b}, \quad c = |c| e^{j\theta_c}$$

el cepstrum complejo es

$$\hat{h}[n] = \frac{2}{n} |b|^{-n} \cos(n\theta_b) u[-n + 1] + \frac{2}{n} |c|^{n} \cos(n\theta_c) u[n - 1]$$

Ejemplos



$$H(z) = \frac{(1 - bz)(1 - b^* z)}{(1 - cz^{-1})(1 - c^* z^{-1})}$$

$$b = 0.97e^{j0.88\pi}, \quad c = 0.97e^{j0.12\pi}$$

$$\hat{h}[n] = \frac{2}{n} 0.97^{-n} \cos(0.88\pi n) u[-n+1] + \frac{2}{n} 0.97^{n} \cos(0.12\pi n) u[n-1]$$

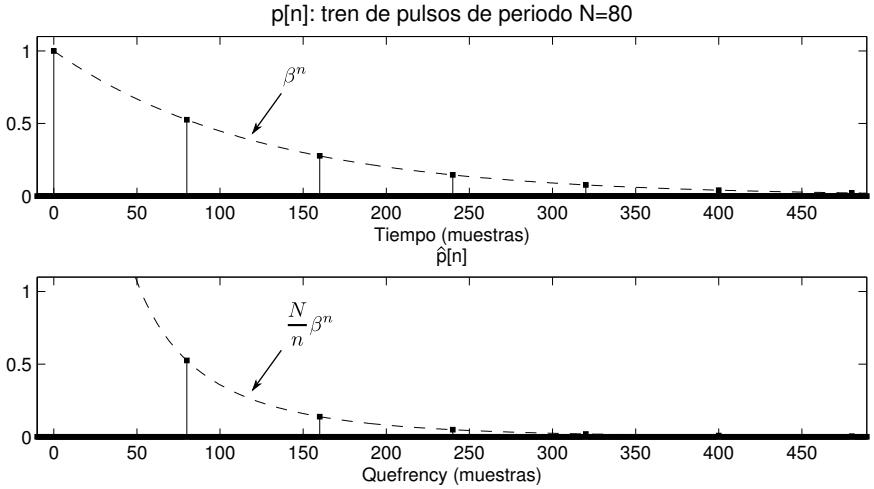
Ejemplos

Tren de pulsos periódico

$$p[n] = \beta^n \sum_{k=0}^{\infty} \delta[n - kN], \text{ con } |\beta| < 1$$

Cepstrum complejo

$$\hat{p}[n] = \frac{N}{n} \beta^n \sum_{k=1}^{\infty} \delta[n - kN]$$



Ejemplos

Tren de pulsos periódico filtrado

$$s[n] = p[n] * h[n],$$

con

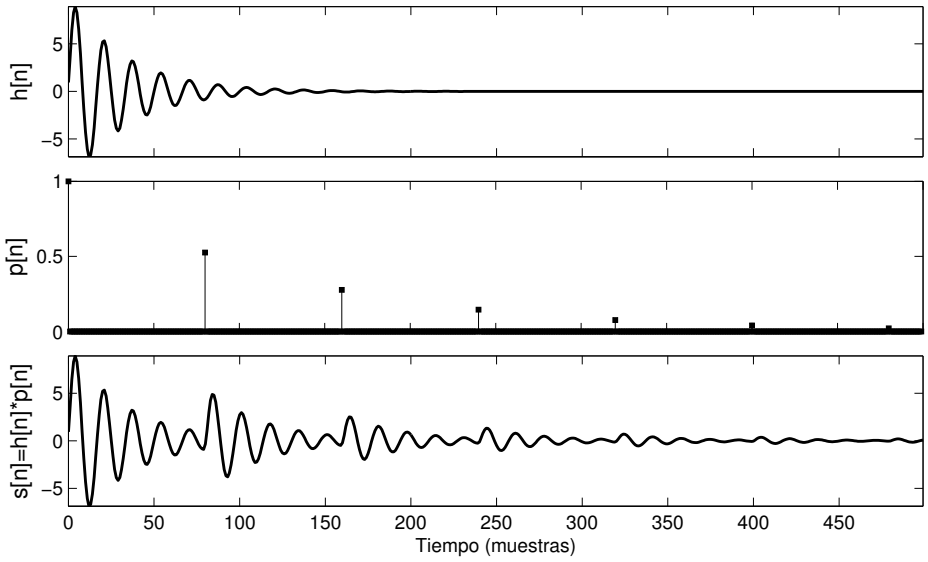
$$H(z) = \frac{(1 - bz)(1 - b^* z)}{(1 - cz^{-1})(1 - c^* z^{-1})}, \quad |b|, |c| < 1$$

$$p[n] = \beta^n \sum_{k=0}^{\infty} \delta[n - kN], \quad |\beta| < 1$$

Forma de onda obtenida con el modelo digital del mecanismo de producción de la voz.

Ejemplos

Tren de pulsos periódico filtrado.

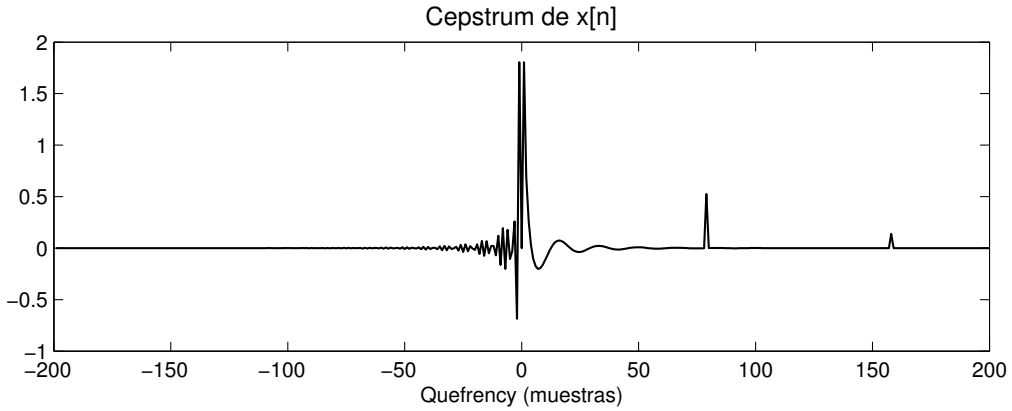


Ejemplos

Cepstrum complejo de tren de pulsos filtrado

$$x[n] = p[n] * h[n] \qquad \hat{x}[n] = \hat{p}[n] + \hat{h}[n]$$

- Componentes de baja quefrency correspondiente a la respuesta del sistema.
- Componentes de alta quefrency correspondiente al tren de impulsos.
- Como en general los componentes no se solapan pueden ser separados.

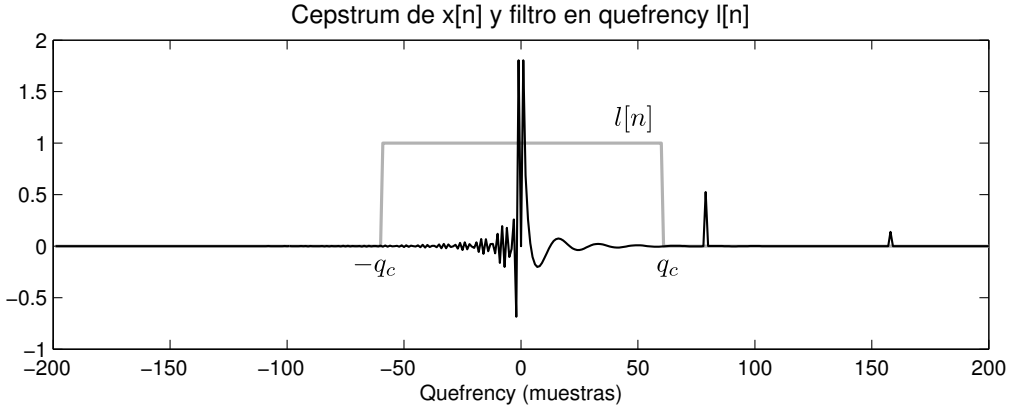


Filtrado homomórfico

Filtrado del cepstrum para deconvolución

$$\hat{y}[n] = \hat{x}[n] l[n] \qquad l[n] = \begin{cases} 1, & |n| \leq q_c \\ 0, & |n| > q_c \end{cases}$$

- q_c: quefrency de corte.
- q_c debe ser ser menor al período del tren de pulsos.



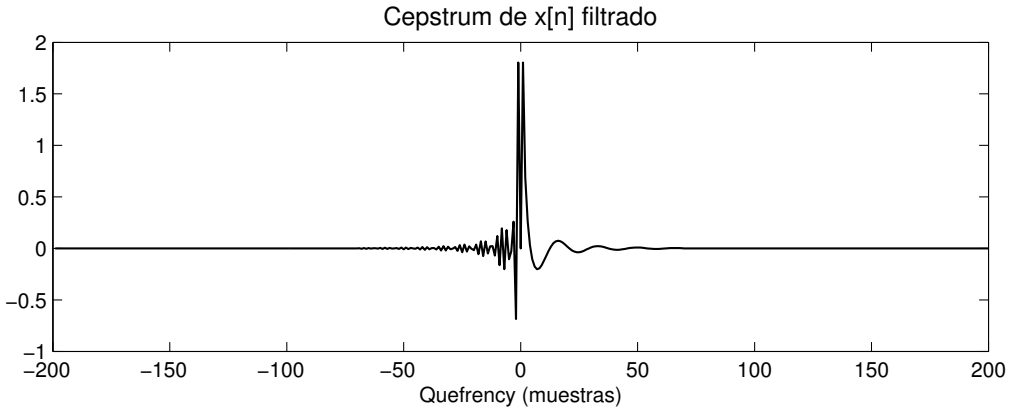
Filtrado homomórfico

Filtrado del cepstrum para deconvolución

Como

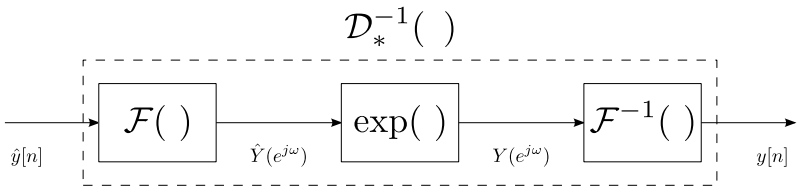
$$\hat{h}[n] \approx 0, \quad |n| > q_c$$
$$\hat{p}[n] = 0, \quad |n| \leq q_c$$

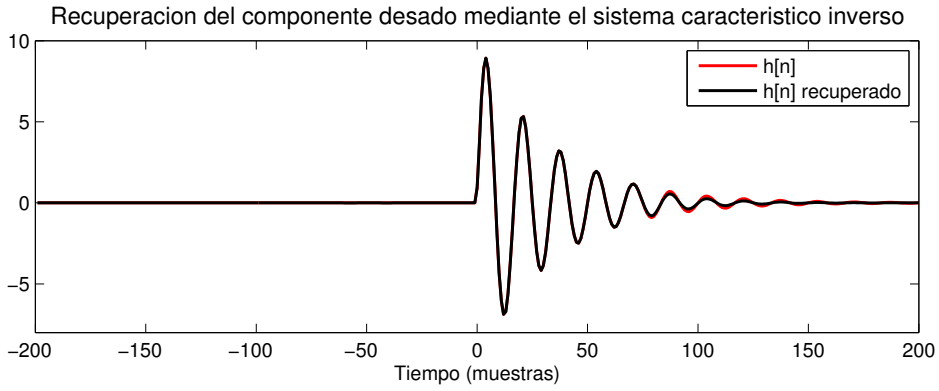
$$\begin{aligned}\hat{y}[n] &= \left(\hat{p}[n] + \hat{h}[n]\right) l[n] \\ &= \hat{p}[n]l[n] + \hat{h}[n]l[n] \\ &\approx \hat{h}[n].\end{aligned}$$



Filtrado homomórfico

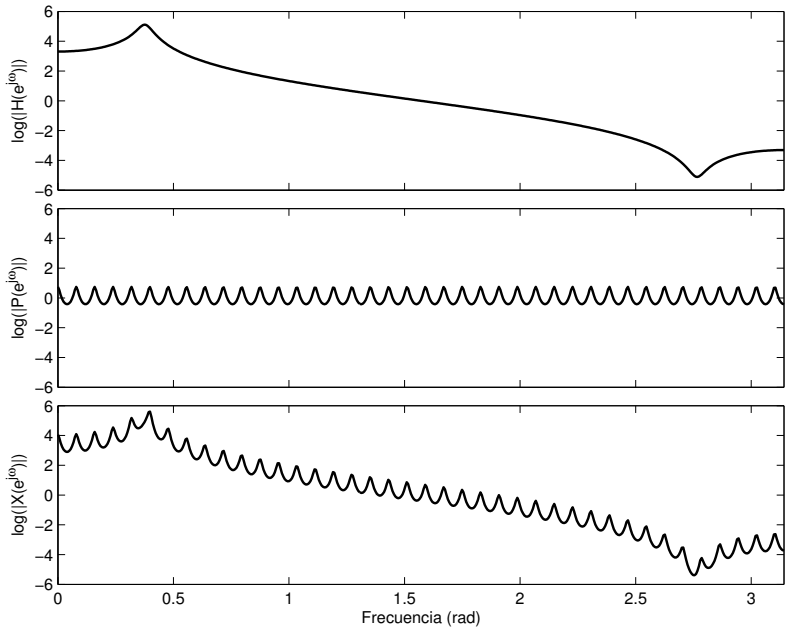
Procesamiento con el sistema característico inverso



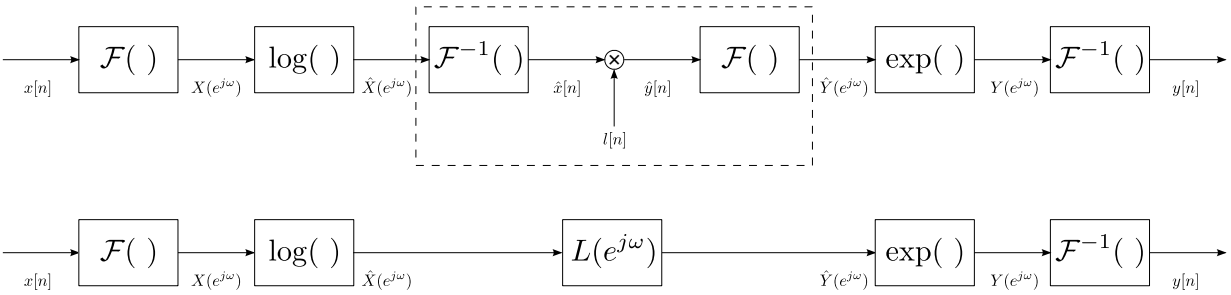


Filtrado homomórfico

Interpretación como suavizado espectral



Filtrado homomórfico



Interpretación como suavizado espectral

- Se considera a $\hat{X}(e^{j\omega}) = \log X(e^{j\omega})$ como una señal en el tiempo con:
  - componentes de baja frecuencia: espectro de la respuesta del sistema
  - componentes de alta frecuencia: espectro del tren de pulsos
- Los componentes se separan con un filtro pasabajos con respuesta al impulso

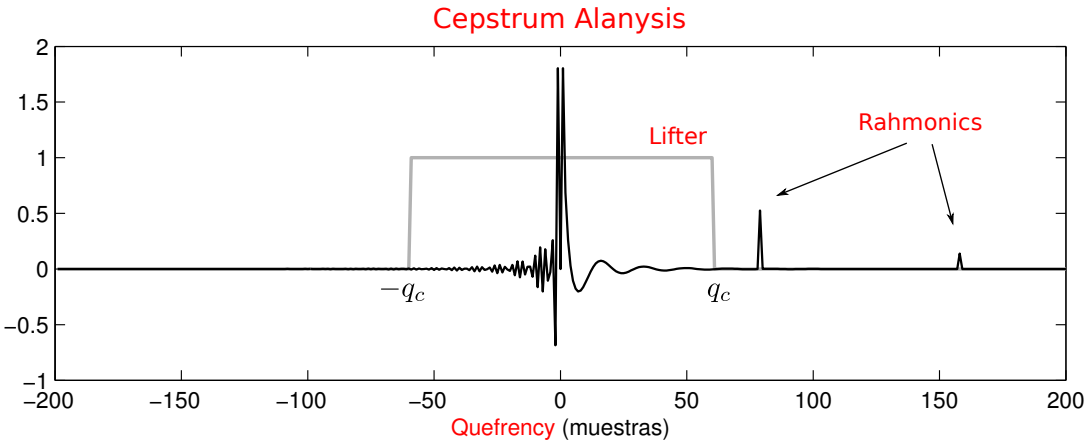
$$L(e^{j\omega}) = \mathcal{F}(l[n])$$

- El espectro suavizado se obtiene como $\hat{Y}(e^{j\omega}) = L(e^{j\omega}) * \log X(e^{j\omega})$

Filtrado homomórfico

Nomenclatura [Bogert et al., 1963]

- El cepstrum $\hat{x}[n]$ puede interpretarse como el espectro de $\hat{X}(e^{j\omega}) = \log X(e^{j\omega})$.
- Se intercambia el dominio del tiempo y la frecuencia.



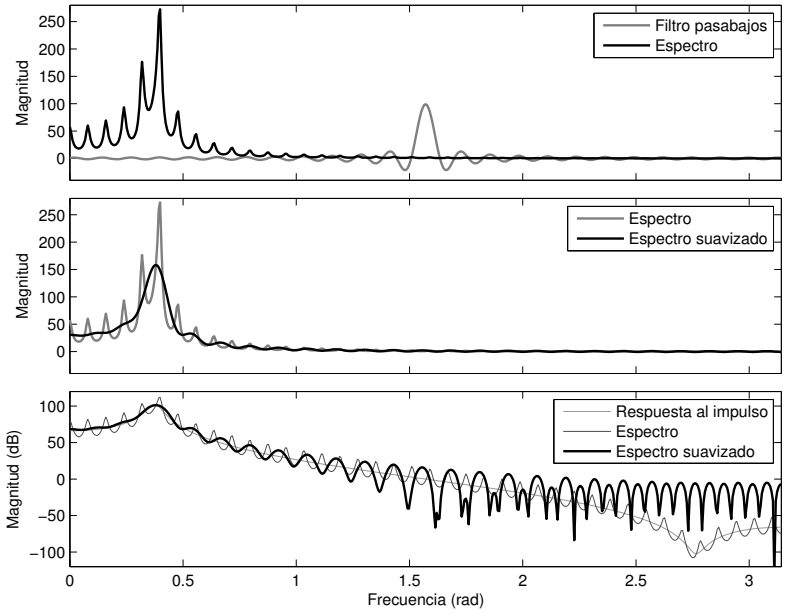
Filtrado homomórfico

Importancia del logaritmo

- Suavizado del espectro en lugar del logaritmo del espectro.
  - El rango dinámico del espectro es grande
  - Al filtrarlo pasabajos, las regiones de alta energía enmascaran a las de baja energía por derramamiento.
  - La envolvente espectral no se aproxima correctamente en las regiones de altas frecuencias.
- El logaritmo comprime el espectro reduciendo su rango dinámico.

Filtrado homomórfico

Importancia del logaritmo



Cepstrum complejo discreto

Cálculo usando la DFT

- En la práctica se trabaja con señales de largo finito.
- Para calcular el cepstrum de una secuencia $x[n]$ de largo N se emplea la DFT de N muestras

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}$$

$$\hat{X}[k] = \log X[k] = \log(|X[k]|) + j\angle X[k]$$

$$\hat{x}_N[n] = \frac{1}{N} \sum_{n=0}^{N-1} \hat{X}[k] e^{j\frac{2\pi}{N}kn}$$

- Dificultades computacionales:
  - aliasing
  - desenvolvimiento de fase a partir de las muestras

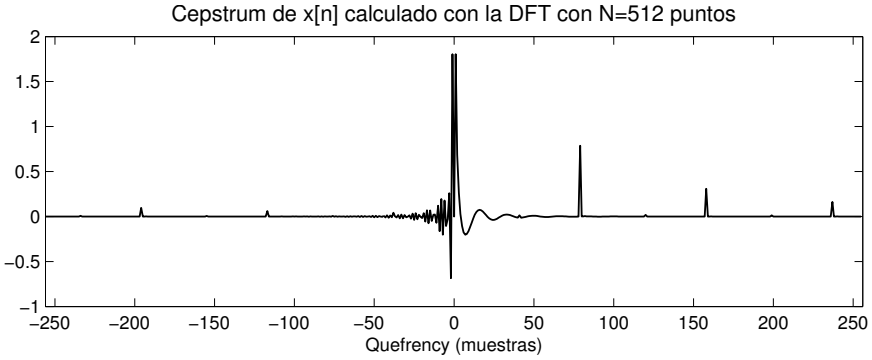
Cepstrum complejo discreto

Aliasing

- $\hat{x}[n]$ es de duración infinita $\Rightarrow$ $\hat{x}_N[n]$ es una versión con aliasing de $\hat{x}[n]$:

$$\hat{x}_N[n] = \sum_{r=-\infty}^{+\infty} \hat{x}[n - rN]$$

- Se evita empleando la DFT con largo suficientemente grande (ej.: N=1024 muestras).



Efecto del enventanado

Enventanado

- Se multiplica la señal con ventana suavizante $w[n]$ (ej. ventana de Hann)

$$s[n] = w[n](p[n] * h[n])$$

- El enventanado destruye el modelo convolucional.
- Si $w[n]$ es suave respecto a $h[n]$, el modelo convolucional sigue siendo válido,

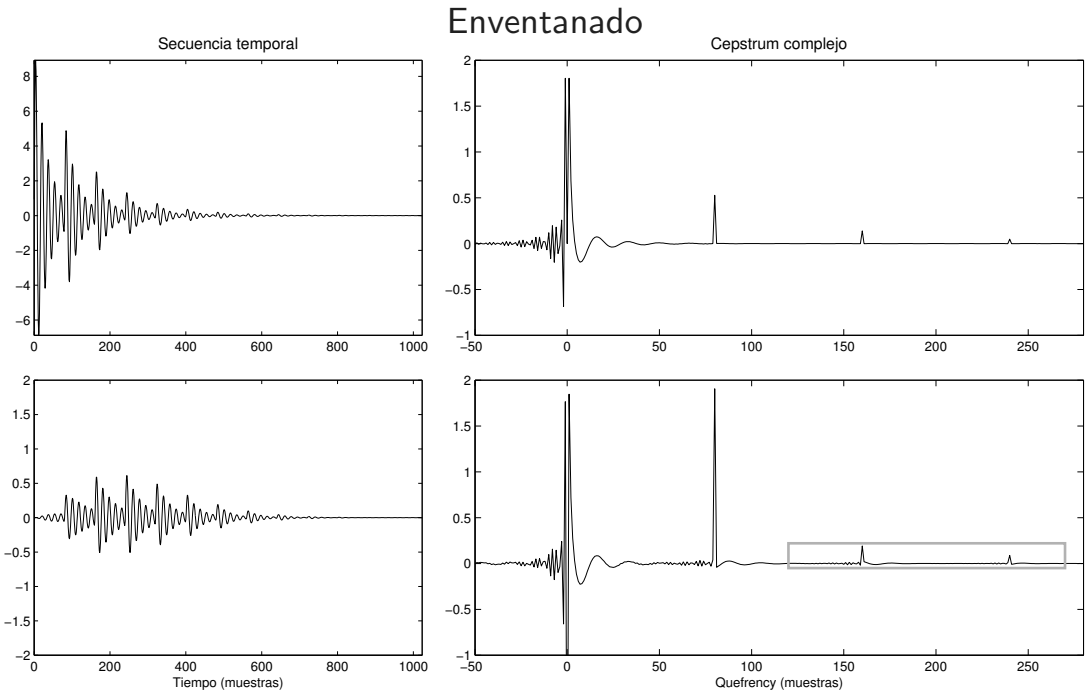
$$s[n] \approx (w[n]p[n]) * h[n])$$

- El cepstrum de la señal enventanada es

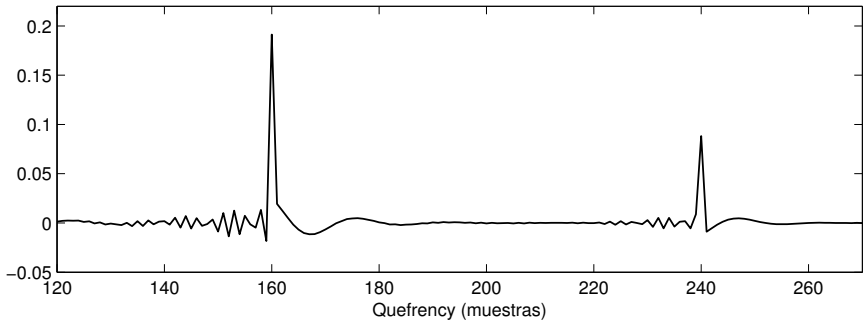
$$\hat{s}[n] = \hat{p}[n] + D[n] \sum_{k=-\infty}^{+\infty} \hat{h}[n - kP]$$

  - $\hat{p}[n]$ es el cepstrum de $w[n]p[n]$
  - $D[n]$ es una función en forma de campana centrada en $n = 0$ que depende de $w[n]$
  - P es el período de la señal

Efecto del enventanado



Efecto del enventanado



Aliasing de $\hat{h}[n]$

- Aparecen copias del cepstrum de $h[n]$ en múltiplos del período.
- Se produce aliasing entre las copias de $\hat{h}[n]$.
- Filtrado homomórfico: Para atenuar la distorsión provocada por el aliasing la quefrency de corte debe ser elegida tal que

$$q_c \leq \frac{P}{2}$$

Efecto del enventanado

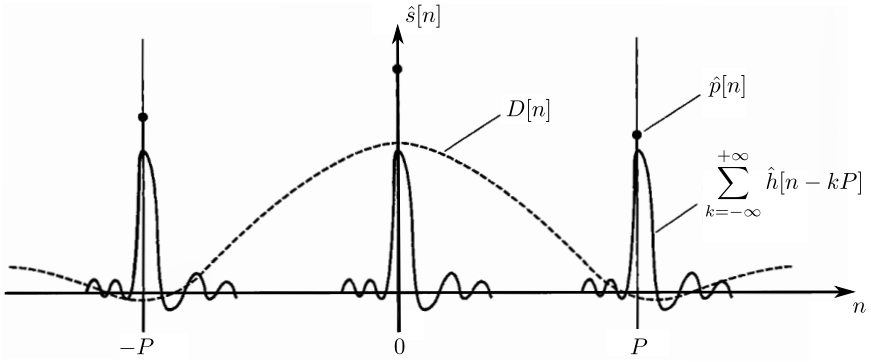
Largo óptimo de ventana

- Cepstrum de señal enventanada

$$\hat{s}[n] = \hat{p}[n] + D[n] \sum_{k=-\infty}^{+\infty} \hat{h}[n - kP]$$

  - $D[n]$ es una función en forma de campana centrada en $n = 0$ que depende de $w[n]$.
  - $D[n]$ decae mas lentamente cuanto mayor es el largo de $w[n]$.

[Quatieri, 2002]



Efecto del enventanado

Largo óptimo de ventana

- La ventana $w[n]$ debe ser lo suficientemente larga para que $D[n]$ no distorsione a $\hat{h}[n]$.
- La ventana $w[n]$ debe ser lo suficientemente corta para que $D[n]$ atenúe la copias de $\hat{h}[n]$

Para ventanas típicas (Hann, Hamming), un largo de entre 2 a 4 períodos de la señal balancean el compromiso.

Ejemplo

- Una señal de periodo de 200 Hz muestreada a 44100 Hz tiene un periodo de 220 muestras.
- El tamaño de ventana es mas pequeño que el empleado habitualmente para realizar un análisis espectral.

Cepstrum de sonidos sonoros

Modelo

- Transferencia del modelo desde la fuente glotal hasta la salida a través de los labios

$$H(z) = A_v G(z) V(z) R(z)$$

  - A_v: ganancia de la fuente glotal
  - $G(z)$: modelo del pulso glotal
  - $R(z)$: modelo de radiación
  - $V(z)$: transferencia del tracto vocal

- La salida en el dominio del tiempo es

$$x[n] = p[n] * h[n],$$

  donde $p[n]$ es un tren de pulsos periódico ideal.
- Se busca separar $h[n]$ y $p[n]$ a partir de $x[n]$ y el modelo.

Cepstrum de sonidos sonoros

Procesamiento

1. Se enventana la señal con una ventana suavizante.
2. Se calcula el cepstrum real o complejo usando la DFT.

$$\hat{x}_N[n] = \hat{p}_N[n] + \hat{h}_N[n]$$

   - $\hat{p}_N[n]$: picos en múltiplos del período P
   - $\hat{h}_N[n]$: soporte en bajas quefrencys

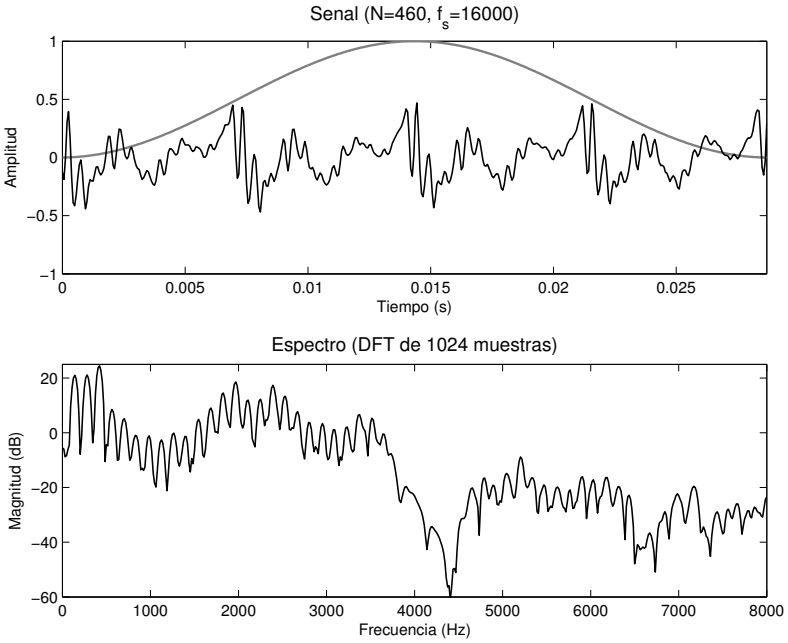
3. Se aplica un lifter para seleccionar la región de baja quefrency de frecuencia de corte

$$q_c \approx \frac{P}{2}.$$

   Se obtiene $\hat{h}_N[n]$.
4. Se aplica el sistema característico inverso al cepstrum liftrado para obtener la respuesta al impulso $h[n]$.

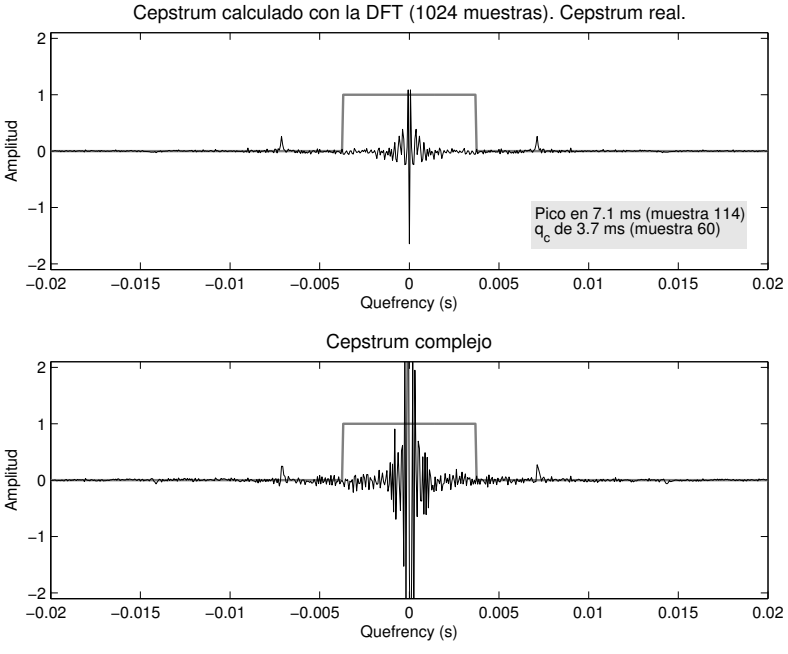
Ejemplo

Enventanado



Ejemplo

Cepstrum real y complejo



Ejemplo

Cálculo del cepstrum de secuencia de fase mínima

En el caso de secuencias de fase mínima, el cepstrum complejo puede obtenerse a partir del liftrado del cepstrum.

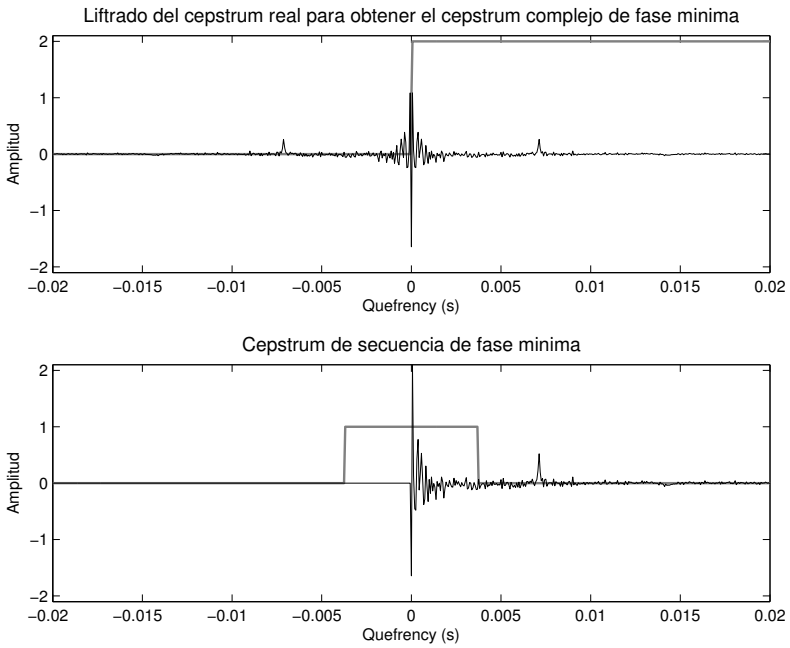
$$\hat{x}[n] = \begin{cases} 0 & n < 0 \\ c[n] & n = 0 \\ 2c[n] & n > 0 \end{cases} \Rightarrow \hat{x}[n] = c[n]l[n], \quad \text{con } l[n] = \begin{cases} 0 & n < 0 \\ 1 & n = 0 \\ 2 & n > 0 \end{cases}$$

Posibilidades

- Cepstrum
- Cepstrum complejo
- Cepstrum de secuencia de fase mínima a partir del cepstrum
- Cepstrum de secuencia de fase máxima a partir del cepstrum

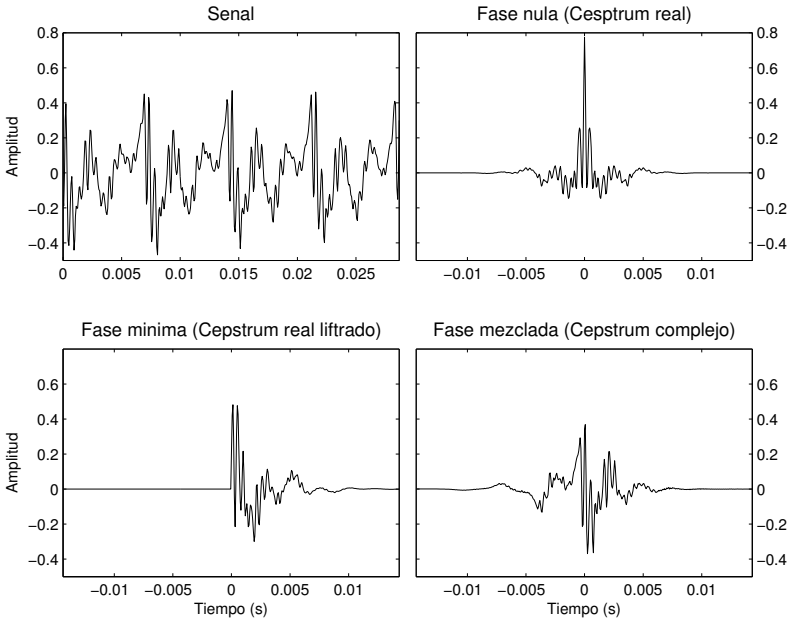
Ejemplo

Cálculo del cepstrum de secuencia de fase mínima



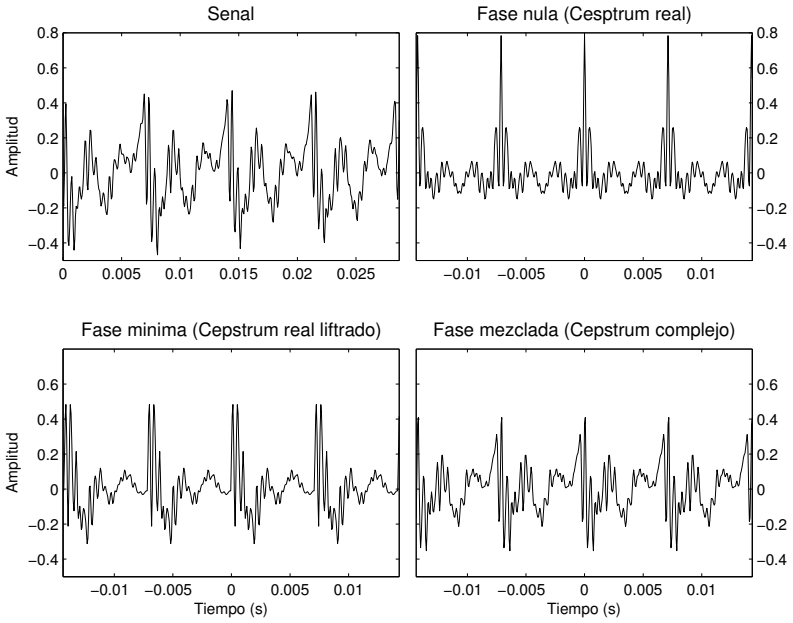
Ejemplo

Respuesta al impulso



Ejemplo

Tren de impulsos filtrado con la respuesta al impulso



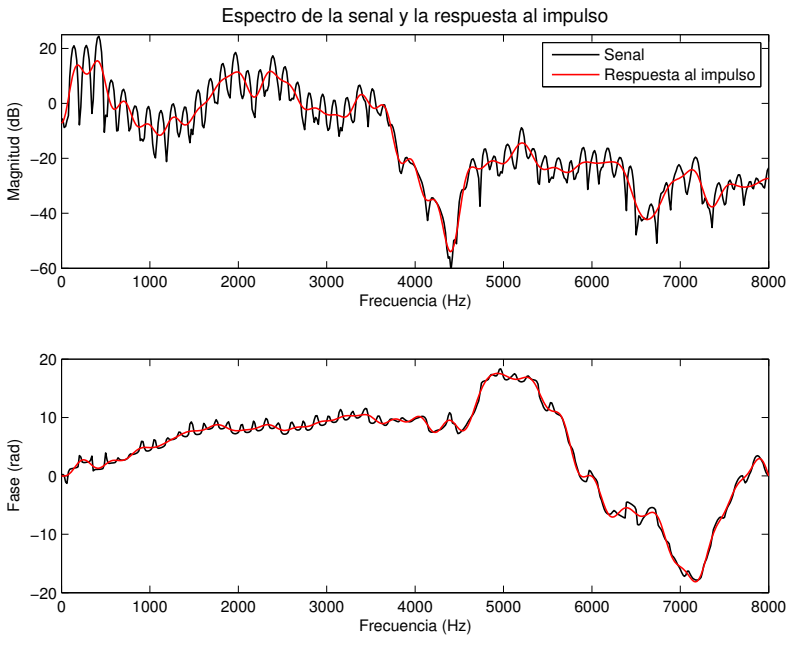
Ejemplo

Características de la respuesta al impulso

- Cepstrum real: respuesta al impulso de fase nula
  - Simétrica respecto al origen
  - La energía se concentra en el origen
  - La forma de onda es “picuda”
- Cepstrum real liftrado: respuesta al impulso de fase mínima
  - Secuencia hacia adelante
  - La energía se concentra en el origen (en menor medida)
  - Si la respuesta del tracto vocal es de fase mínima, se obtiene una estimación correcta. En caso contrario se obtiene una aproximación cruda de la fase de la señal.
- Cepstrum complejo: respuesta al impulso de fase mezclada
  - La estimación es una buena aproximación de la respuesta al impulso del tracto vocal.
  - El cepstrum complejo es mas difícil de calcular debido a la dificultad de manipular la fase de señales de audio.

Ejemplo

Espectro de la respuesta al impulso. Suavizado espectral.

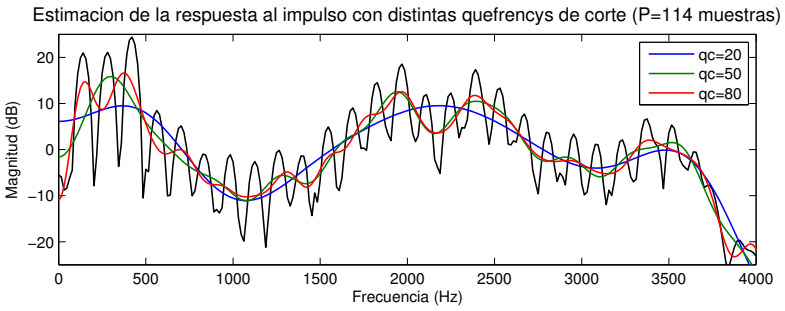


Ejemplo

Espectro de la respuesta al impulso. Suavizado espectral.

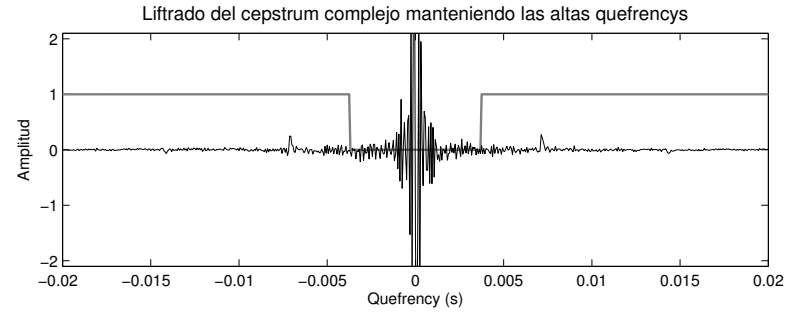
Observaciones

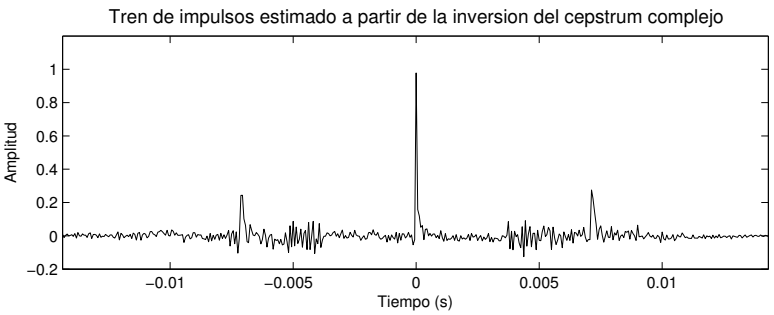
- Puede estimarse la frecuencia de las formantes a partir de los picos del espectro de la respuesta al impulso.
- El estimador espectral suavizado no pasa por los picos espectrales. No es un estimador de la envolvente espectral.
- El grado de suavizado depende de la quefrency de corte al liftrar el cepstrum.



Ejemplo

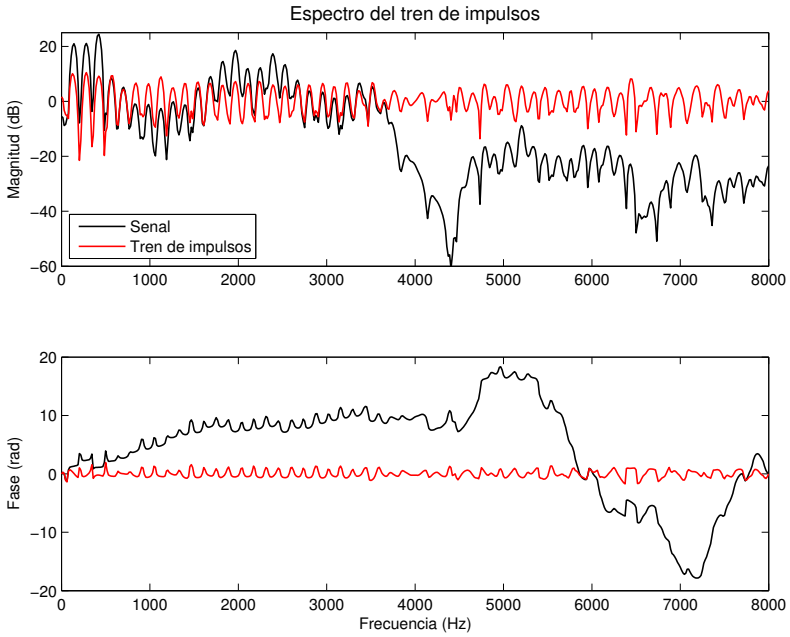
Estimación del tren de impulsos enventanado.





Ejemplo

Estimación del tren de impulsos enventanado.



Ejemplo

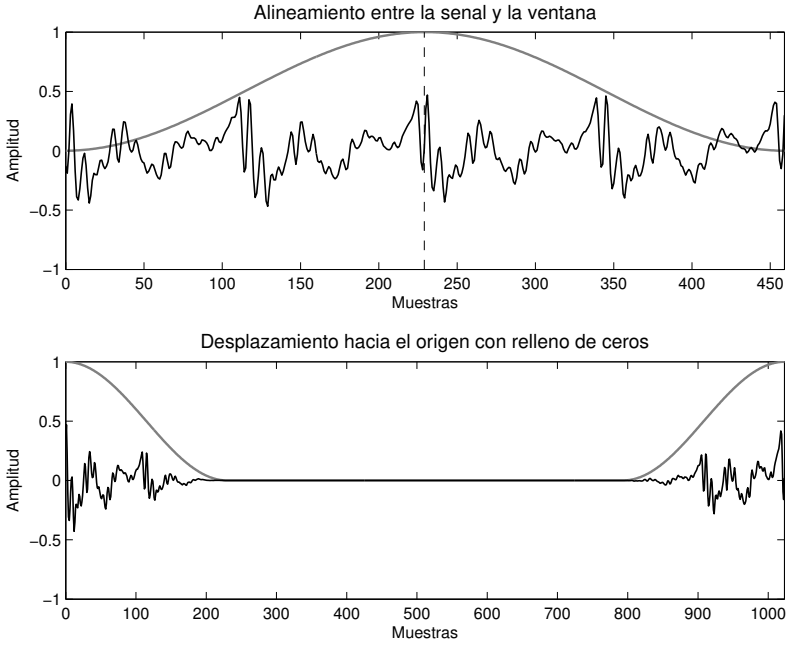
Consideraciones prácticas para el cálculo del cepstrum complejo

- La ventana tiene que tener un número entero de períodos de la señal. El tamaño de la ventana debe ser adaptado al período de la señal.
  - El tamaño de ventana puede obtenerse a partir del pico del cepstrum.
- Centro de la ventana debe estar alineado con la respuesta al impulso.
  - Se busca el máximo del segmento de tiempo corto y se selecciona el primer cruce por cero anterior.
- El centro de la ventana debe ser desplazado al origen ($n = 0$) para eliminar el componente de fase lineal (*fftshift*).

En el caso en que solo interese estimar el espectro de la respuesta al impulso conviene emplear el cepstrum real.

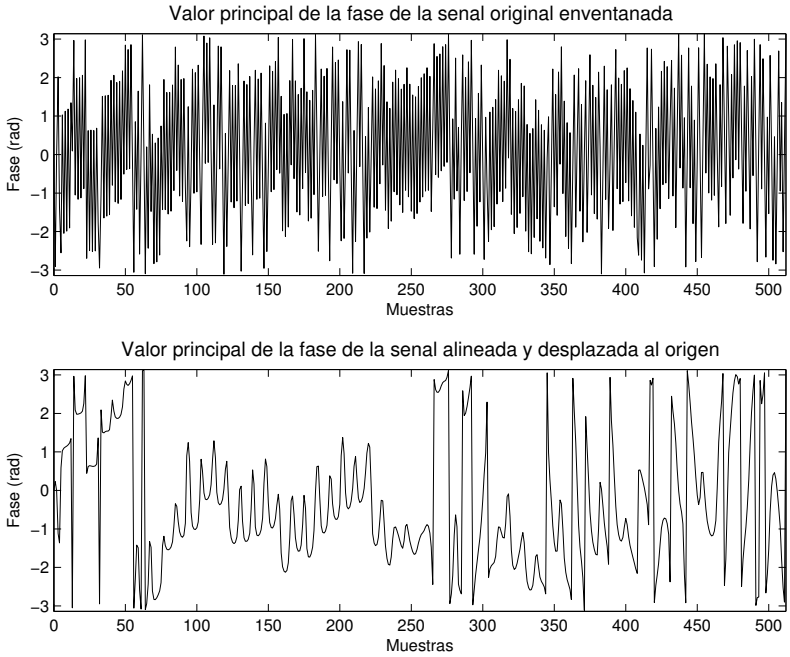
Ejemplo

Consideraciones prácticas para el cálculo del cepstrum complejo.



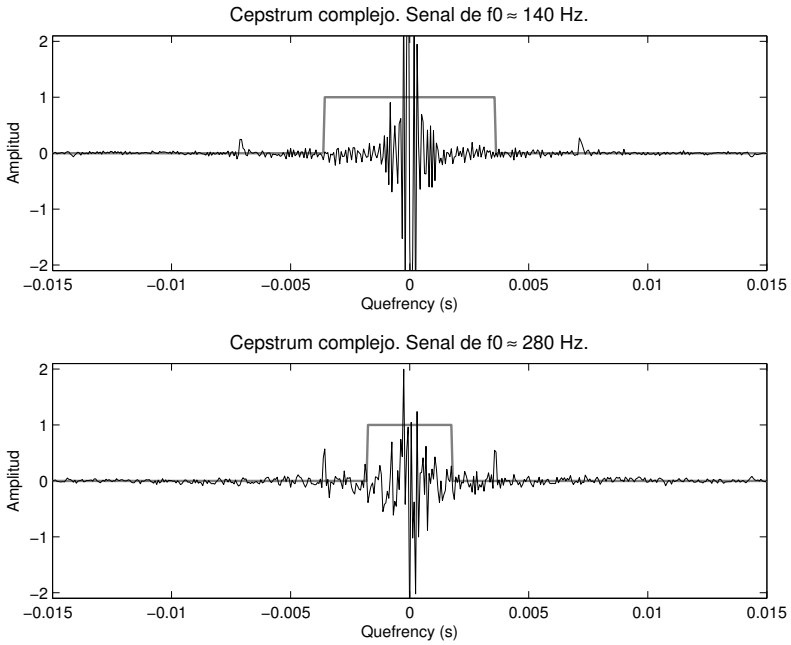
Ejemplo

Consideraciones prácticas para el cálculo del cepstrum complejo.



Ejemplo

Dificultad con sonidos agudos.



Ejemplo

Dificultad con sonidos agudos

1. El primer pico correspondiente a $\hat{p}[n]$ está cerca del origen. El cepstrum de la respuesta al impulso y del tren de pulsos pueden solaparse en la quefrency.
2. Hay aliasing entre las copias de $\hat{h}[n - kP]$ por ser P pequeño.

Filtrado homomórfico

- Como $q_c \leq \frac{P}{2}$, el largo del lifter decrece con el período P.
- El suavizado espectral es mas fuerte en el caso de sonidos agudos.
- El incremento artificial del ancho de banda de las formantes debido al suavizado excesivo, resulta en sonidos mas "apagados" en la síntesis.

El analisis cepstral es mas benevolente con sonidos graves (ej.: voz masculina) que con sonidos agudos (ej.: voz femenina o de niño).

Cepstrum de sonidos sordos

Modelo

- Transferencia del modelo desde la fuente hasta la salida a través de los labios

$$H(z) = A_n V(z) R(z)$$

  - A_n: ganancia de la fuente
  - $R(z)$: modelo de radiación
  - $V(z)$: transferencia del tracto vocal

- La salida en el dominio del tiempo es

$$x[n] = u[n] * h[n],$$

  donde $u[n]$ es ruido blanco que representa la turbulencia en alguna restricción en el tracto vocal.

Cepstrum de sonidos sordos

Procesamiento

1. Se enventana la señal con una ventana suavizante.
2. Se calcula el cepstrum real usando la DFT.

$$\hat{x}_N[n] = \hat{u}_N[n] + \hat{h}_N[n]$$

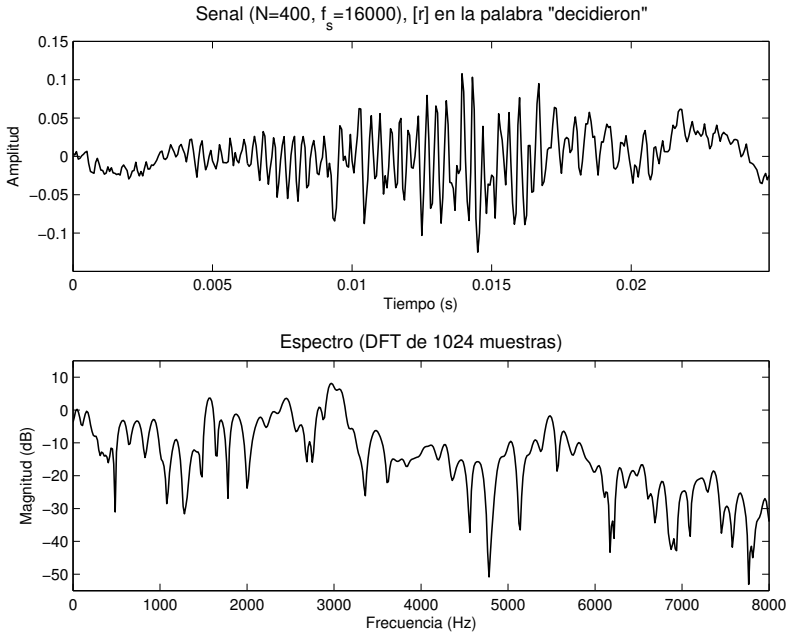
  - $\hat{u}_N[n]$: componentes en todas las quefrencys
  - $\hat{h}_N[n]$: componentes en bajas quefrencys

Observaciones

- Los dos componentes se solapan en las quefrencys y no pueden ser separados.
- La interpretación del filtrado homomórfico como suavizado espectral hace que la técnica sea aplicable.
  - Las fluctuaciones rápidas se deben a la excitación.
  - La envolvente espectral corresponde a la respuesta al impulso.

Ejemplo

Análisis de sonidos sordos.



Cepstrum de sonidos sordos

Observaciones

- Es difícil desenvolver la fase del espectro
  - La fase de secuencias aleatorias salta arbitrariamente entre muestras en la frecuencia discreta.
- La fase de la función de transferencia no es perceptivamente signicativa para sonidos sordos.

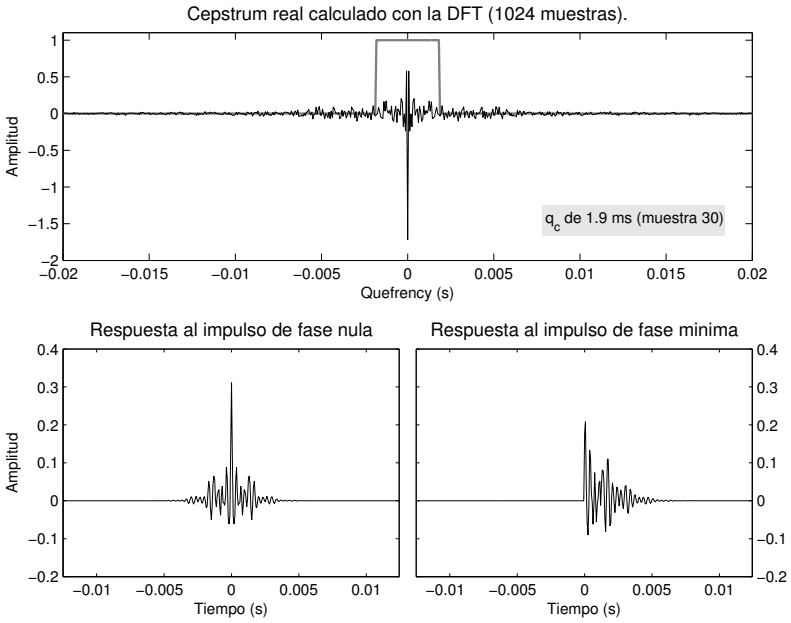
En la práctica se emplea el cepstrum real para procesar sonidos sordos.

Procesamiento

3. Se aplica un lifter para seleccionar la región de bajas quefrencys. Se obtiene $\hat{h}_N[n]$.
4. Se aplica el sistema característico inverso al cepstrum liftrado para obtener la respuesta al impulso $h[n]$.

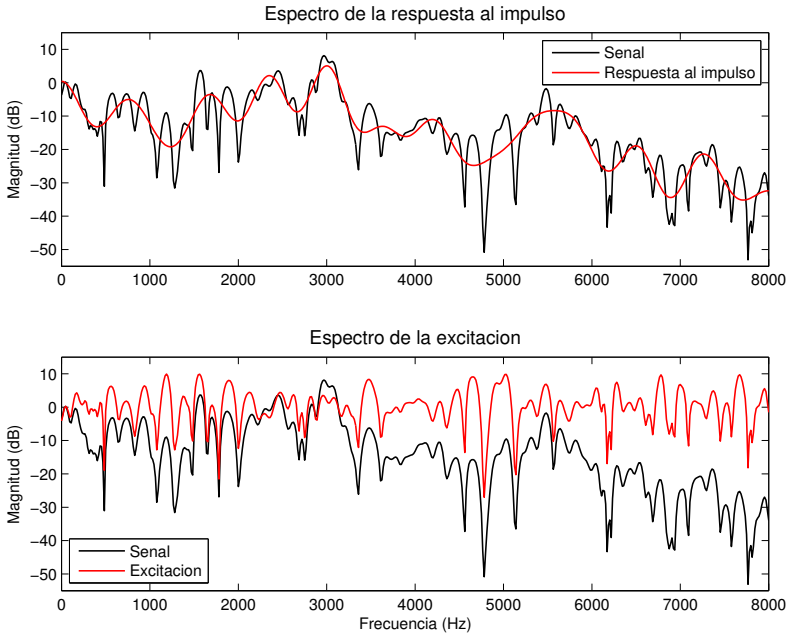
Ejemplo

Cepstrum real y respuestas al impulso.



Ejemplo

Estimación de la respuesta al impulso y la excitación.



Detección de frecuencia fundamental

Altura y sonoridad

- La presencia de un pico es un indicador de la sonoridad del sonido.
  - Presencia de pico: sonido sonoro
  - Ausencia de pico: sonido sordo
- La posición del pico indica el período de la señal.

Algoritmo [Noll, 1967]: si el pico supera cierto umbral, se establece que el sonido es sonoro, y la posición del pico indica el período de la señal

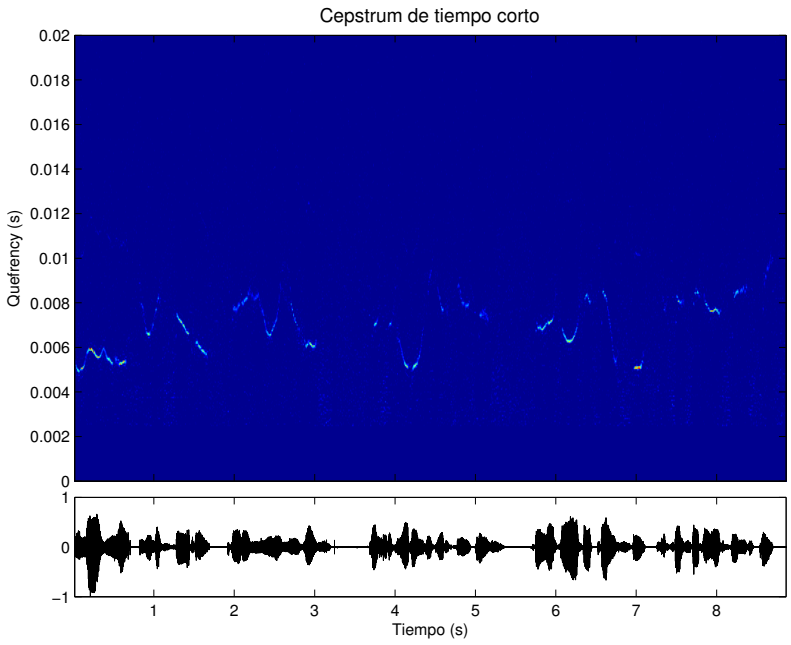
Ejemplo

Procesamiento de señal de voz masculina

- $f_s = 16000$ Hz.
- $N = 600$ muestras (≈ 38 ms)
- $N_{DFT} = 1024$ muestras
- Salto = N/4 muestras (≈ 10 ms)

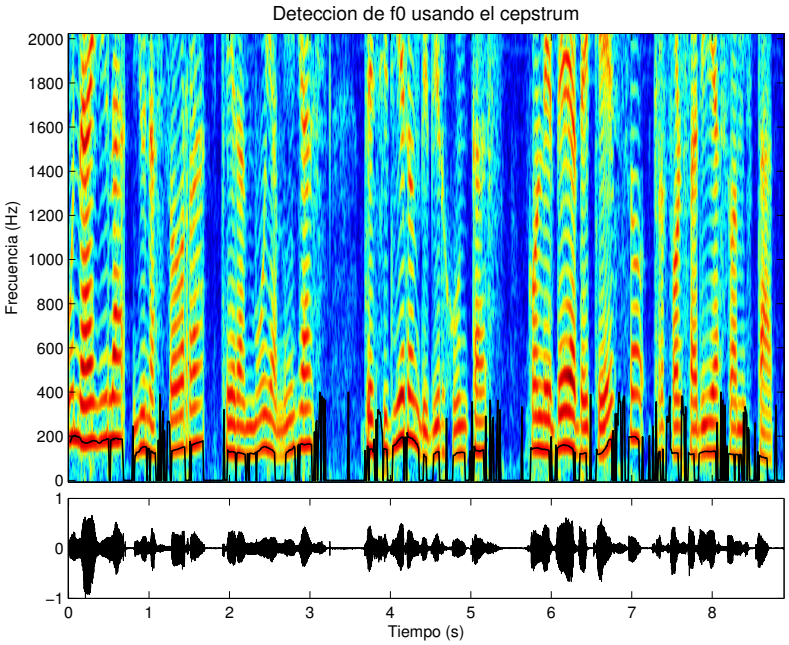
Detección de frecuencia fundamental

Cepstrum de tiempo corto.



Detección de frecuencia fundamental

Estimación obtenida a partir del pico del cepstrum.



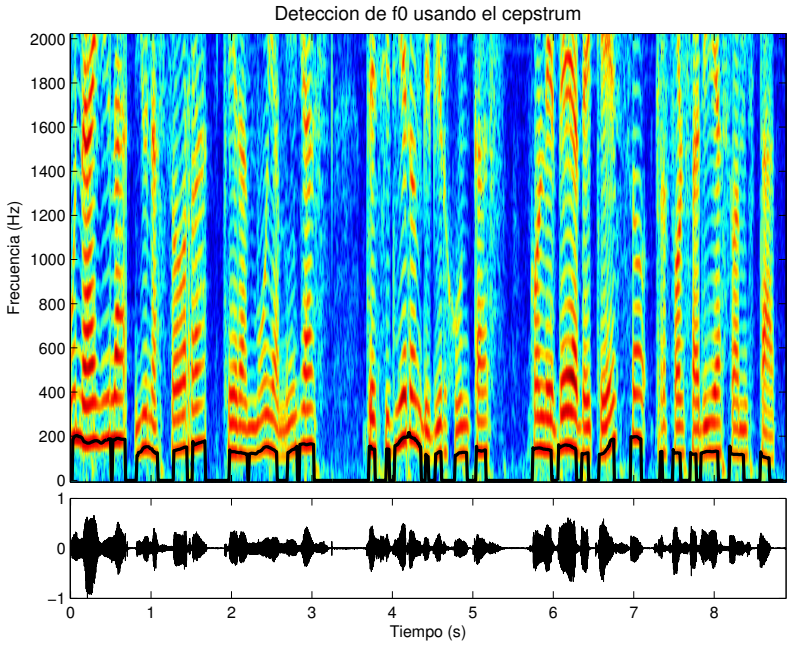
Detección de frecuencia fundamental

Dificultades

- La ausencia de pico o la presencia de un pico pequeño no es necesariamente una indicación de sonido sordo. La magnitud del pico depende de
  - Cantidad de períodos que entran en la ventana.
  - Posición relativa de la señal en la ventana.
- Desviación del modelo en caso de señales de banda limitada
  - Si el espectro logarítmico no tiene oscilaciones, el cepstrum no tiene picos.
- En general, el pico del cepstrum ocurre en quefrencys donde los componentes de la respuesta en frecuencia están muy atenuados.
  - Puede usarse un umbral pequeño para detectar el pico
  - En los casos en donde el cepstrum falla en indicar claramente la sonoridad se puede complementar con información adicional, como la tasa de cruces por cero y la energía.

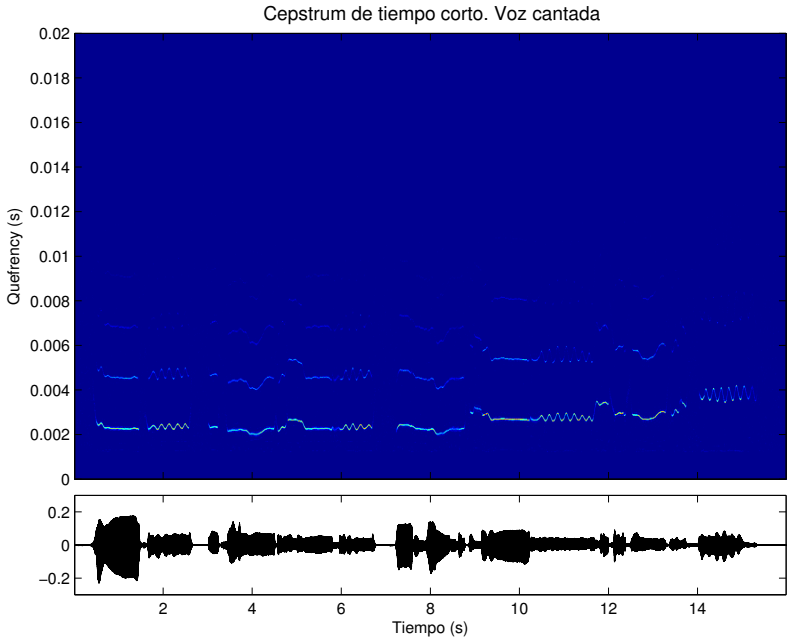
Detección de frecuencia fundamental

Estimación postprocesada.



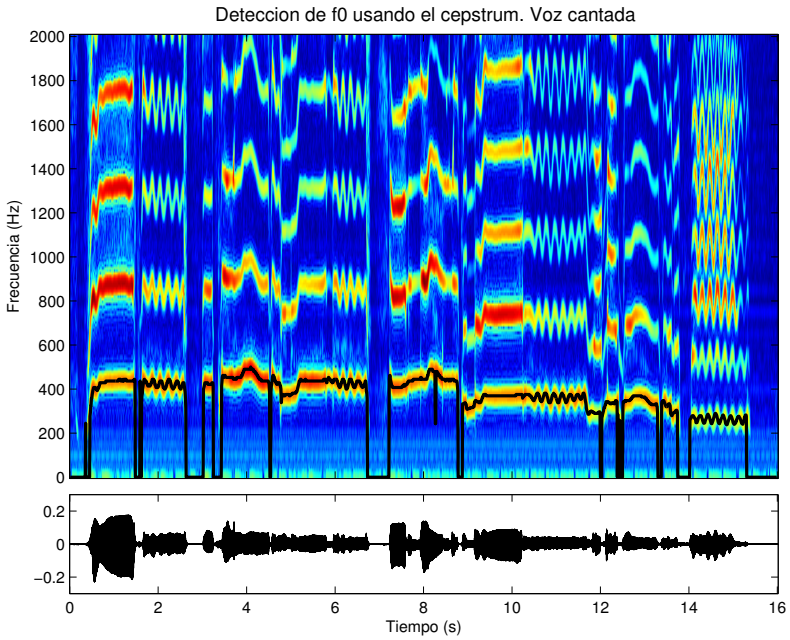
Detección de frecuencia fundamental

Cepstrum de tiempo corto. Voz cantada.



Detección de frecuencia fundamental

Estimación obtenida a partir del pico del cepstrum. Voz cantada.



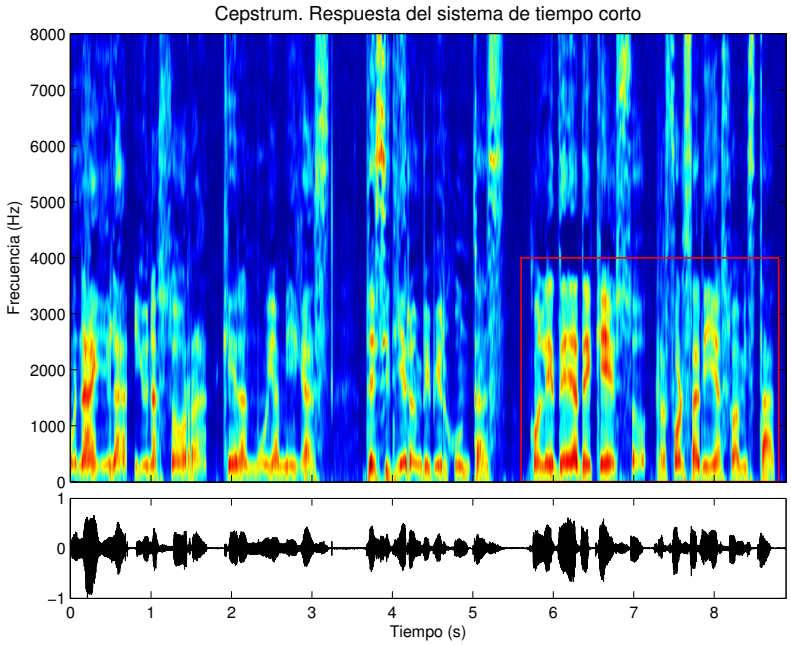
Detección de formantes

Espectro suavizado

- A partir del cepstrum liftrado puede obtenerse el espectro suavizado, que es una estimación de la transferencia del tracto vocal.
- El espectro suavizado muestra la estructura de resonancias del segmento de audio.
- La frecuencia de las formantes puede estimarse a partir de los picos del espectro suavizado.

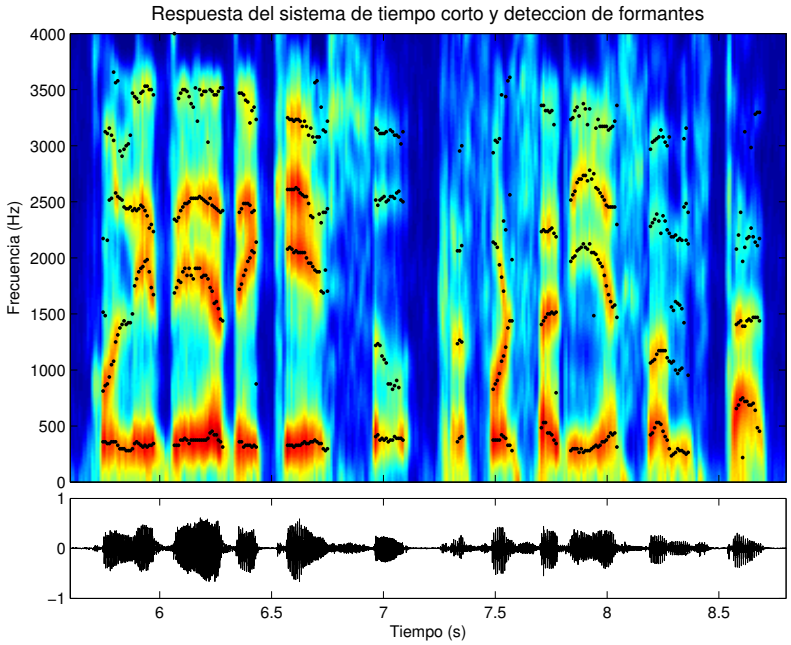
Detección de formantes

Función de transferencia de tiempo corto.



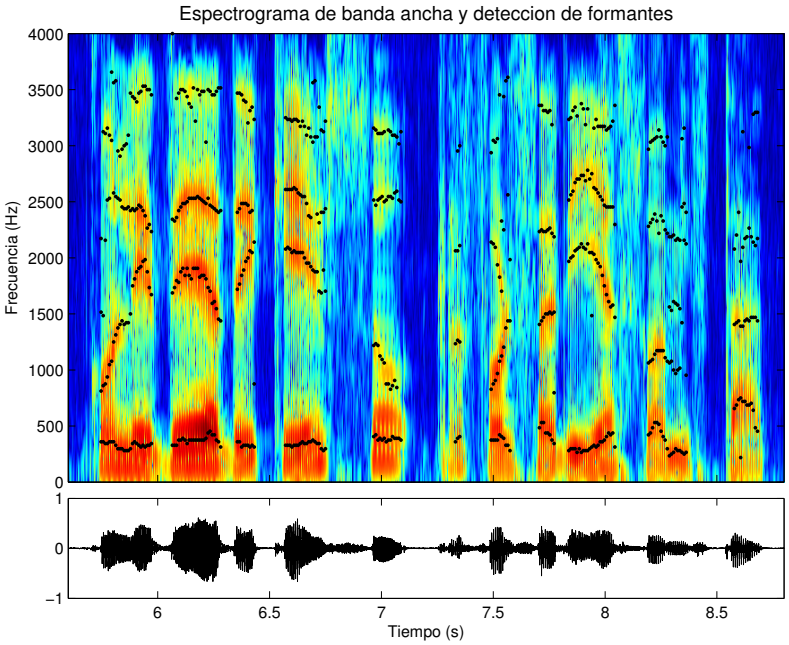
Detección de formantes

Función de transferencia de tiempo corto y máximos locales principales.



Detección de formantes

Máximos mas prominentes y espectrograma de banda ancha.



Codificación de voz

Estructuras de análisis y síntesis

- La eliminación de las altas quefrencys del cepstrum conduce a la respuesta al impulso del tracto vocal.
- La eliminación de las bajas quefrencys del cepstrum conduce a la excitación.
- La convolución entre la respuesta al impulso y la excitación reconstruye exactamente a la señal original de tiempo corto.
- Con una estructura de solapamiento y suma se reconstruye exactamente la señal original completa.

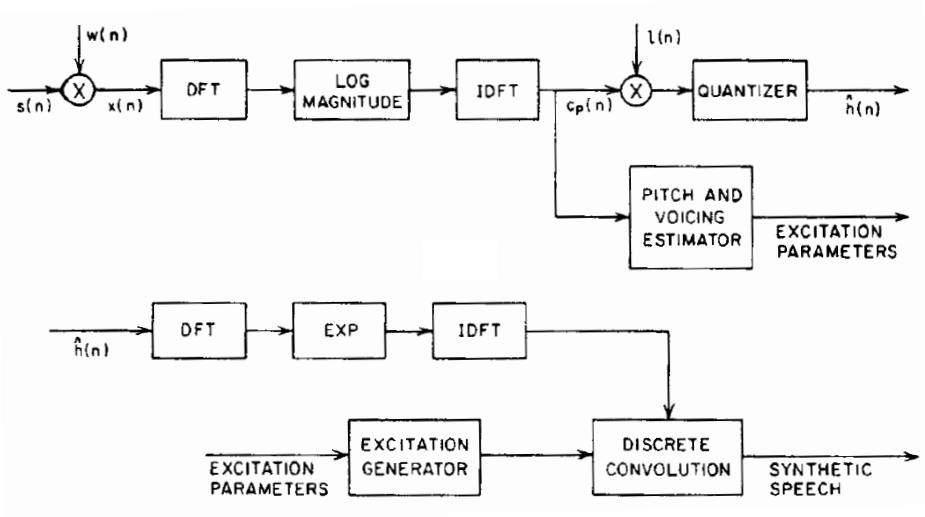
Codificación de voz

Vocoder homomórfico [Oppenheim, 1969]

- Mediante la representación compacta obtenida con el cepstrum es posible sintetizar de forma eficiente una estimación de la señal.
- Etapa de análisis
  - Cálculo del cepstrum *real* de la señal de tiempo corto.
  - Estimación de la sonoridad y la frecuencia fundamental a partir del cepstrum.
  - Representación de la señal: sonoridad, frecuencia fundamental y primeros coeficientes del cepstrum ($\hat{h}[n]$).
- Etapa de síntesis
  - Inversión del cepstrum $\hat{h}[n]$ para obtener $h[n]$.
  - Síntesis de la excitación a partir de la sonoridad y la frecuencia fundamental.
  - Síntesis de la señal de tiempo corto con la convolución de la excitación con la respuesta al impulso $h[n]$.
  - Solapamiento y suma de las reconstrucciones de tiempo corto.

Codificación de voz

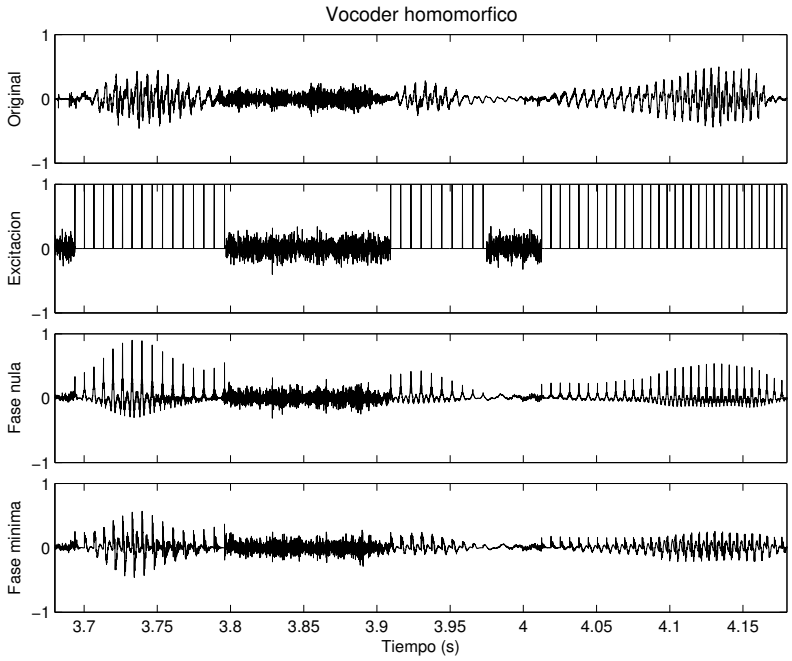
Esquema del vocoder homomórfico



[Rabiner and Schafer, 1978]

Codificación de voz

Síntesis con el vocoder homomórfico



Codificación de voz

Ejemplo: codificación de voz a baja tasa de bits

Procesamiento:

- $N = 512$ muestras
- Salto $= 256$ muestras

Codificación:

- 30 componentes del cepstrum
- Estimación de $f0$

Cada 256 muestras se transmiten 31: tasa de compresión $\approx$ 8

Observaciones

- Al emplear el cepstrum real, solo es posible reconstruir señales de fase nula o fase mínima.
- Se produce deterioro de la calidad de la síntesis.
- Agregando complejidad al sistema es posible usar el cepstrum complejo y reconstruir también la fase [Quatieri, 1979],
  - Tamaño de ventana adaptivo al período de la señal
  - Alineamiento entre la señal y la ventana

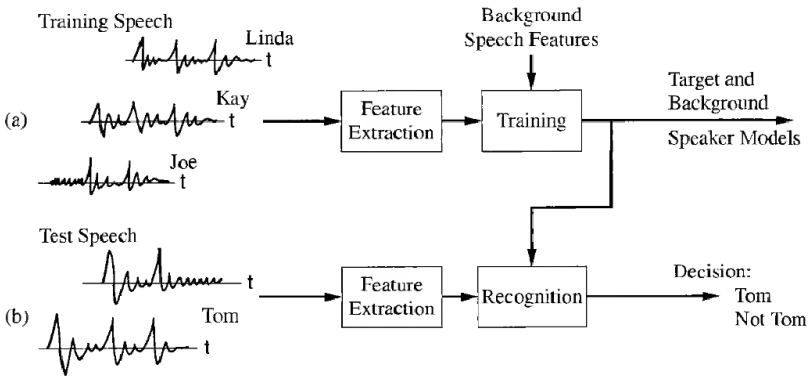
Clasificación de señales de audio

Problema

Identificar la categoría a la que pertenece una nueva observación en base a un conjunto de entrenamiento de observaciones cuya categoría es conocida.

- identificación del hablante
- voz, musica, ruido
- artista
- genero musical

[Quatieri, 2002]



Clasificación de señales de audio

Ejemplo: Verificación del hablante

- Entrenamiento
  - Extracción de características de las señales de voz.
  - Construcción de base de datos con un modelo de cada hablante.
- Reconocimiento:
  - Extracción de características de la señal de voz de prueba.
  - Medida de distancia entre las características y los modelos en la base de datos.
  - Identificación en función de la distancia.

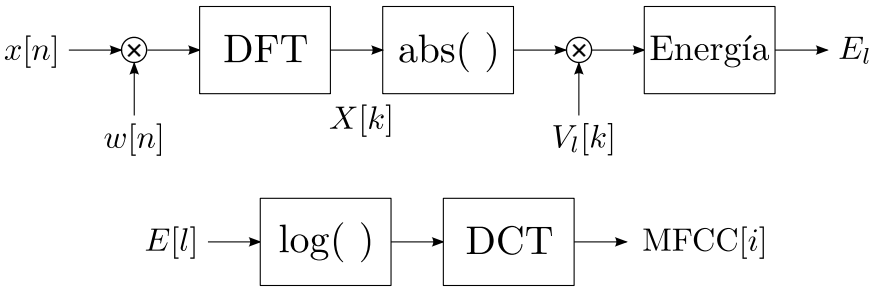
MFCC [Davis and Mermelstein, 1980]

- Características efectivas:
  - envolvente de la STFT
  - basadas en modelos del sistema auditivo
- Coeficientes cepstrales en la escala mel (MFCC)
  - combina ambas características
  - explota la propiedad de decorrelación del cepstrum

Clasificación de señales de audio

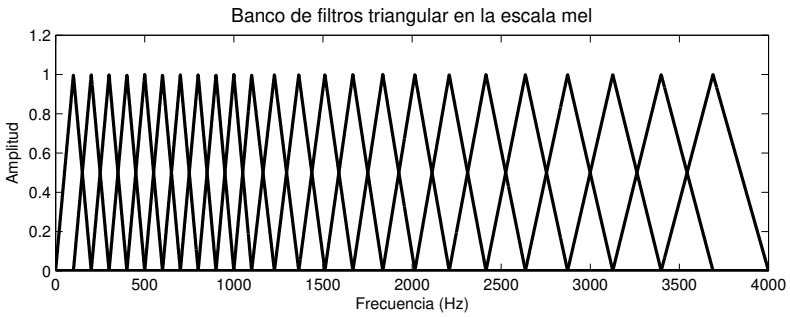
Cálculo de los MFCC

- Cálculo de la DFT de la señal enventanda
- Multiplicación con banco de filtros inspirado en el banco de filtros auditivo
- Cálculo de la energía en cada filtro.
- Cálculo de la DCT de la energía logarítmica como si fuera una señal.



Clasificación de señales de audio

Cálculo de los MFCC



Banco de filtros

Filtro triangulares con ancho de banda

- constante hasta 1000 Hz.
- constante en la escala mel sobre 1000 Hz.

Clasificación de señales de audio

Cálculo de los MFCC

- Cálculo de la energía en el filtro l-ésimo:

$$E[l] = \sum_k |V_l[k]X[k]|^2, \quad \text{para } l = 0, \ldots, L-1,$$

donde L es el número de filtros del banco.

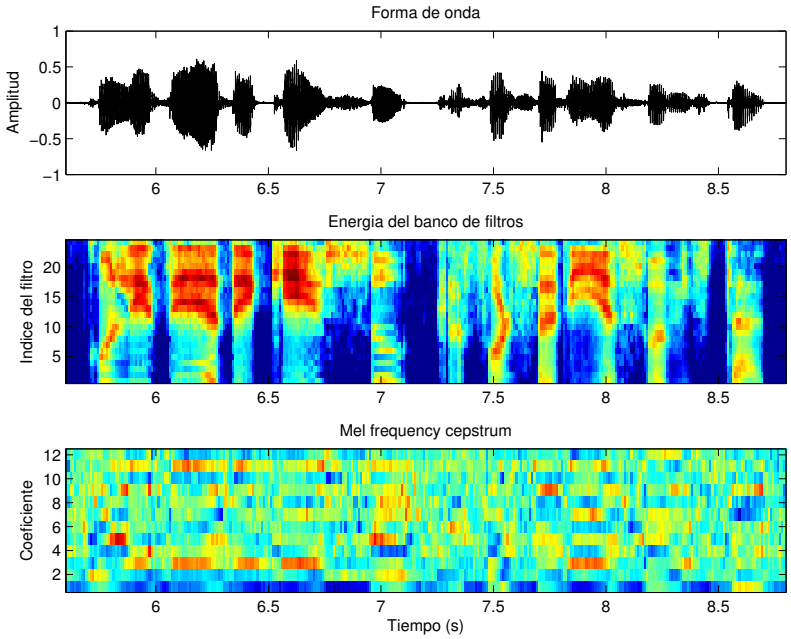
- Cálculo de los coeficientes cepstrales: transformada inversa de Fourier de la energía logarítmica explotando la propiedad de función par del cepstrum real:

$$MFCC[i] = \sum_{l=0}^{L-1} \log(E[l]) \cos\left(\frac{2\pi}{L} li\right), \quad \text{con } i = 0, \ldots, L-1.$$

- Usualmente se emplean los MFCC liftrados.

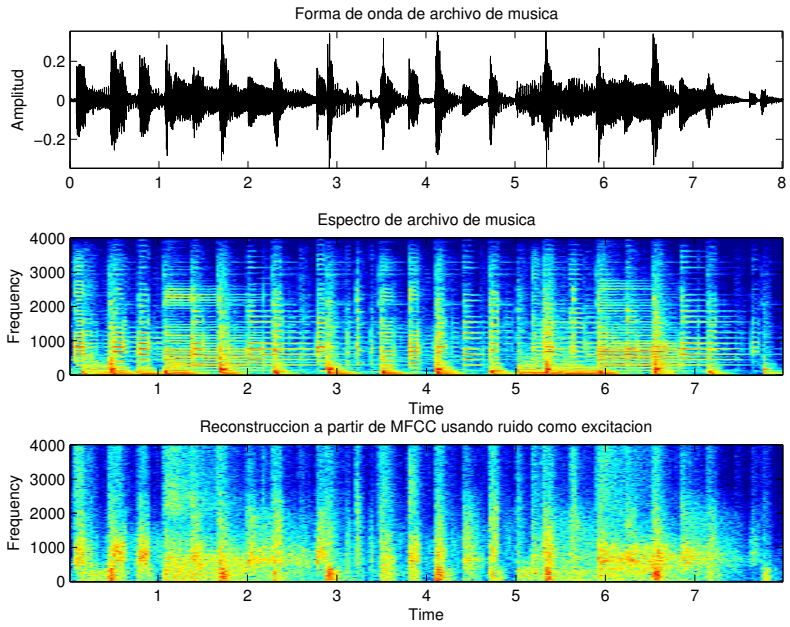
Clasificación de señales de audio

Ejemplo de MFCC de señal de voz



Clasificación de señales de audio

Ejemplo de MFCC de señal de música



Referencias I

Bogert, B. P., Healy, M. J. R., and Tukey, J. W. (1963).
The quefrency alanysis of times series for echos: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking.
In *Proc. of the Symposium on Time Series Analysis*, pages 209–243. Wiley.

Davis, S. and Mermelstein, P. (1980).
Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.
IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4):357–366.

Noll, A. M. (1967).
Cepstrum pitch determination.
Journal of the Acoustical Society of America, 41(2):293–309.

Referencias II

Oppenheim, A. V. (1965).
Superposition in a class of nonlinear systems.
PhD thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science.

Oppenheim, A. V. (1969).
A speech analysis-synthesis system based on homomorphic filtering.
Journal of the Acoustical Society of America, 45(2):458–465.

Quatieri, T. F. (1979).
Minimum and mixed phase speech analysis-synthesis by adaptive homomorphic deconvolution.
IEEE Transactions on Acoustics, Speech and Signal Processing, 27(4):328–335.

Quatieri, T. F. (2002).
Discrete-Time Speech Signal Processing: Principles and Practice.
Prentice Hall PTR, 1st edition.

Referencias III

Rabiner, L. R. and Schafer, R. W. (1978).
Digital Processing of Speech Signals.
Prentice Hall, us edition.