Procesamiento digital de señales de audio

Análisis de Fourier de tiempo corto

Instituto de Ingeniería Eléctrica
Facultad de Ingeniería

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

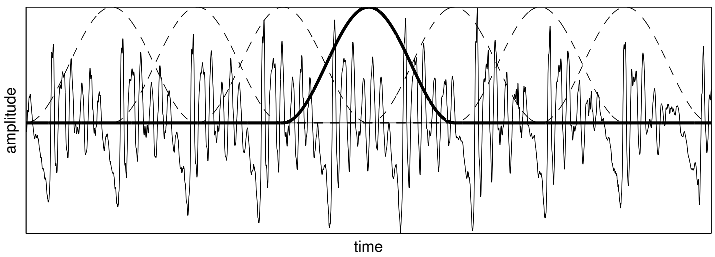
Análisis de Fourier de tiempo corto (STFT)

Short Time Fourier Transform (STFT) [Rabiner and Schafer, 2011]

Representación espectral que refleja variaciones temporales de la señal

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w[n-m]x[m]e^{-j\omega m}$$

- STFT tiempo discreto: $X_n(e^{j\omega})$, n discreta, ω contínua
- STFT discreta: $X_n[k] = X_n(e^{j\omega})|_{\omega=\frac{2\pi}{N}k} \quad k = 0 \dots N-1$



Análisis de Fourier de tiempo corto (STFT)

Short Time Fourier Transform (STFT) [Rabiner and Schafer, 2011]

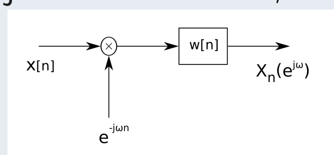
Representación espectral que refleja variaciones temporales de la señal

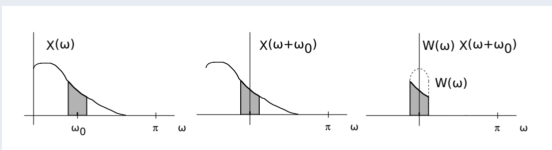
$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w[n-m]x[m]e^{-j\omega m}$$

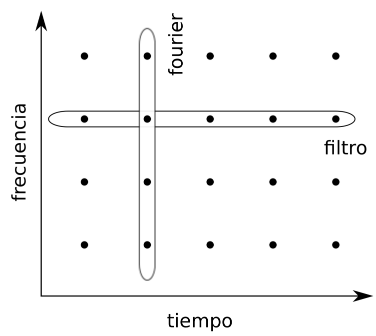
Interpretación

fijo n: TF de $w[n-m]x[m]$

fijo ω: convolución, filtro







Análisis de Fourier de tiempo corto (STFT)

Existencia

$\sum_{n=-\infty}^{\infty} |x[n]| < \infty$ condición suficiente para existencia de TF

$\sum_{n=-\infty}^{\infty} |x[m]w[n-m]| < \infty$ $w[n-m]$ de duración finita ✓

Reconstrucción

$x[m]w[n-m] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X_n(e^{j\omega})e^{j\omega n}d\omega$

si $w[0] \neq 0$, se puede evaluar para $m = n$

$x[n] = \frac{1}{2\pi w[0]}\int_{-\pi}^{\pi} X_n(e^{j\omega})e^{j\omega n}d\omega$

Relación con $R_n(k)$

$S_n(e^{j\omega}) = |X_n(e^{j\omega})|^2 = X_n(e^{j\omega})X_n^*(e^{j\omega})$

$R_n[k] = \sum_{m=-\infty}^{\infty} w[n-m]x[m]w[n-k-m]x[m+k]$

$R_n[k] \overset{\mathrm{TF}}{\longleftrightarrow} S_n[k]$ par de transformadas

Análisis de Fourier de tiempo corto (STFT)

Propiedades de una TF

$X_n(e^{j\omega})$ es una TF respecto a ω:

- periódica de período 2π
- simetría Hermítica para $x[m]w[n-m]$ real
- corrimiento temporal $n_0 \rightarrow e^{-j\omega n_0}X_{n-n_0}(e^{j\omega})$

Efecto del enventanado

$X(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x[m]e^{-j\omega m}, \quad W(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w[m]e^{-j\omega m}$

$w[n-m] \overset{\mathrm{TF}}{\longleftrightarrow} W(e^{-j\omega})e^{-j\omega n}$

convolución de las transformadas de $x[m]$ y $w[n-m]$

$$X_n(e^{j\omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{j\theta})e^{j\theta n}X_n(e^{j(\omega+\theta)})d\theta$$

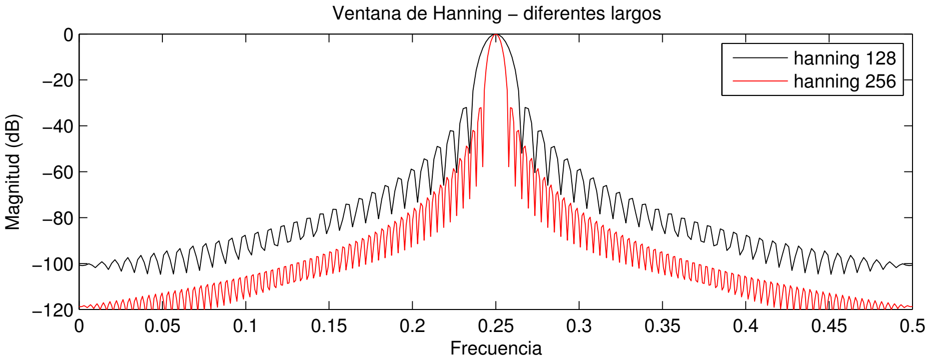
STFT como versión suavizada de la TF de una parte de la señal

Análisis de Fourier de tiempo corto (STFT)

Efecto del enventanado

Propiedades de la ventana:

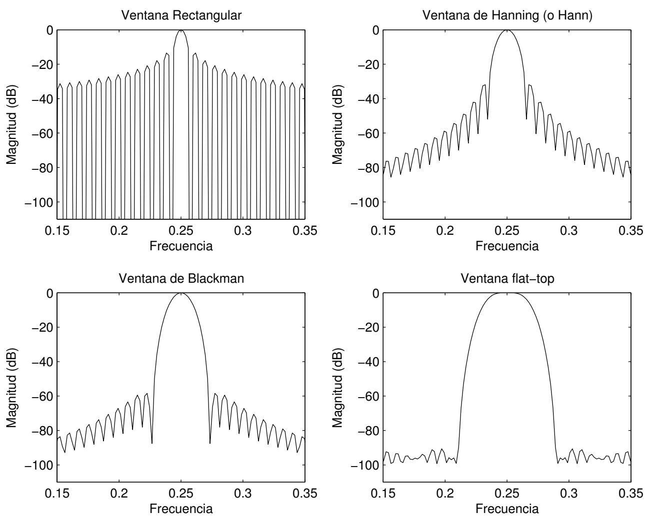
- ancho del lóbulo principal: inversamente proporcional al largo L
- nivel de lóbulos secundarios: independiente del largo
  depende del tipo de ventana
  - rectangular: -13dB, $2\frac{F_s}{L}$ hanning: -31dB, $4\frac{F_s}{L}$



Análisis de Fourier de tiempo corto (STFT)

Efecto del enventanado

- compromiso entre ancho lóbulo principal y nivel lóbulos secundarios
  ejemplo: análisis de frecuencias cercanas (0.1 y 0.15)

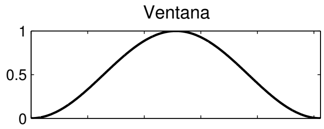


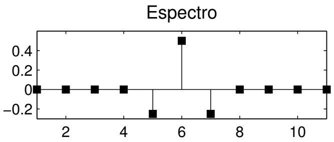
Análisis de Fourier de tiempo corto (STFT)

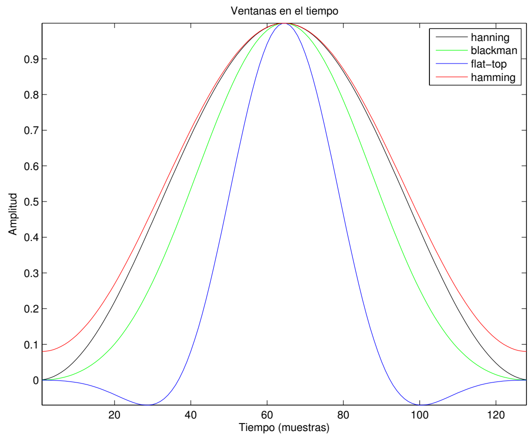
Efecto del enventanado

ventanas típicas: unos pocos componentes en frecuencia no nulos

hann	$a_0 - a_1 \cos(2\pi n/L)$	$a_0 = 0.50,\ a_1 = 0.50$
hamming	$a_0 - a_1 \cos(2\pi n/L)$	$a_0 = 0.54,\ a_1 = 0.46$
blackman	$a_0 - a_1 \cos(2\pi n/L)$ $+a_2 \cos(4\pi/L)$	$a_0 = 0.42,\ a_1 = 0.50$ $a_2 = 0.08$
flat-top	$a_0 - a_1 \cos(2\pi n/L)$ $+a_2 \cos(4\pi/L)$ $-a_3 \cos(6\pi/L)$ $+a_4 \cos(8\pi/L)$	$a_0 \approx 0.22,\ a_1 \approx 0.42$ $a_2 \approx 0.28$ $a_3 \approx 0.08$ $a_4 \approx 0.01$





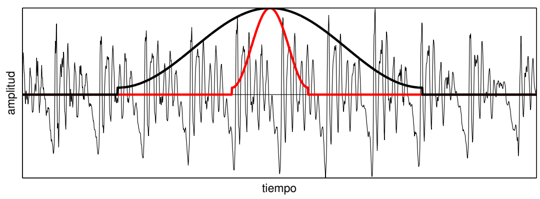


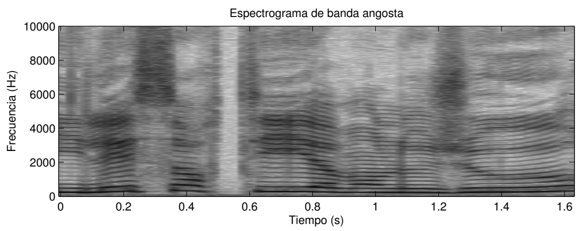
Análisis de Fourier de tiempo corto (STFT)

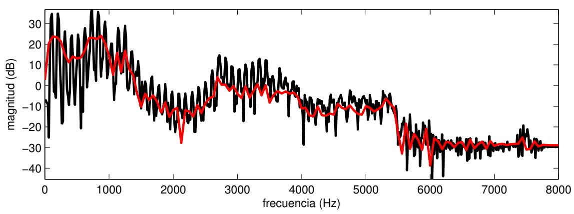
Efecto del enventanado

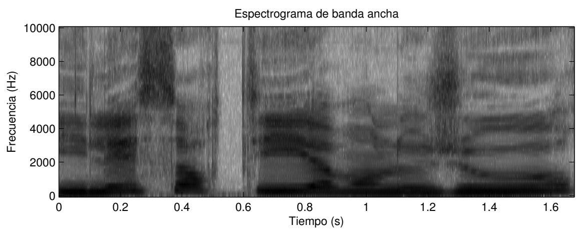
- Análisis de voz usando diferente largo de ventana

64 ms estructura armónica clara
16 ms sólo se distinguen las formantes
pero las formantes pueden cambiar a lo largo de 50 ms





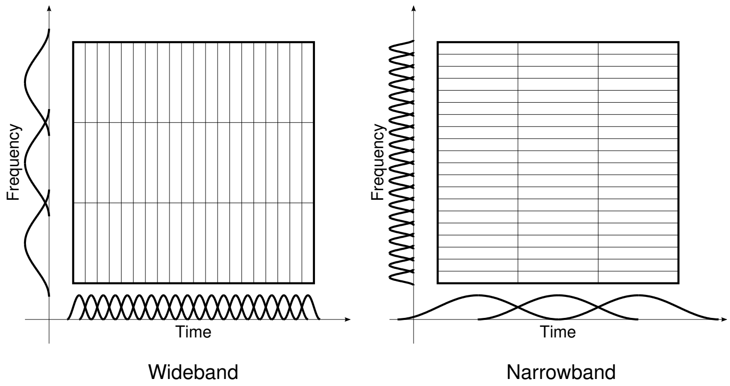




Análisis de Fourier de tiempo corto (STFT)

Espectrograma

- largo L (y forma) de la ventana determina la resolución
  - espectrograma de *banda ancha* (L chico)
  - espectrograma de *banda angosta* (L grande)
- resolución constante en tiempo-frecuencia



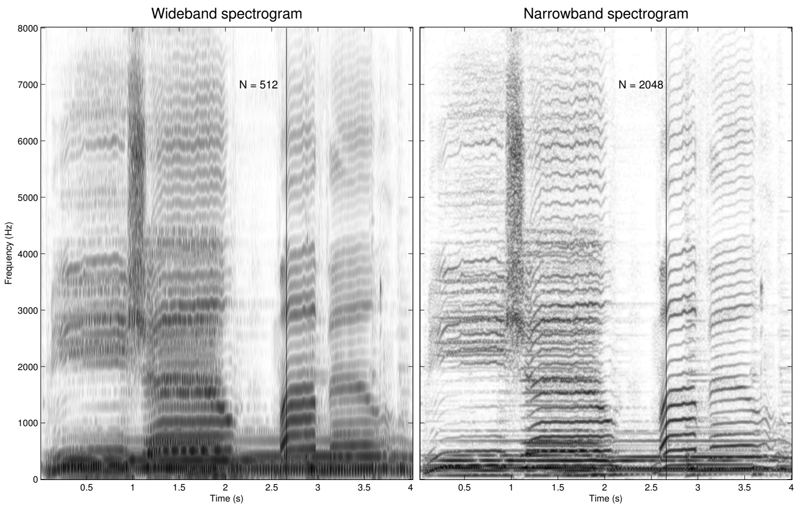
Análisis de Fourier de tiempo corto (STFT)

Banda ancha

- pobre resolución espectral
- buena resolución temporal

Banda angosta

- buena resolución espectral
- pobre resolución temporal



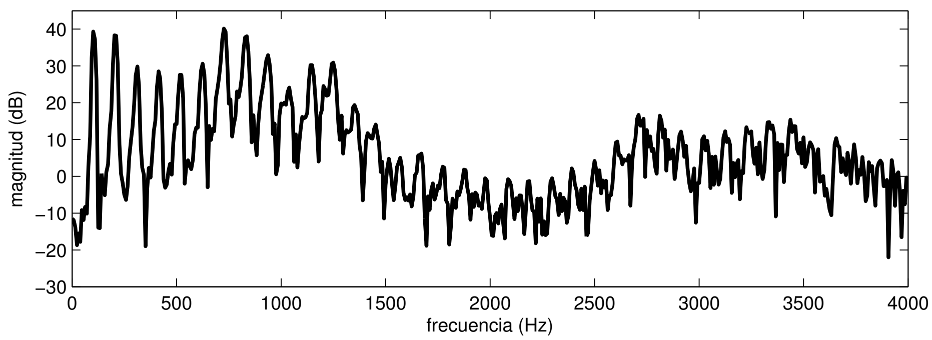
Análisis de Fourier de tiempo corto (STFT)

Tasa de muestreo de la STFT [Rabiner and Schafer, 2011]

- necesidad de muestrear en tiempo y frecuencia produciendo una representación sin *alias* de la cual se pueda reconstruír la señal
- tasa de muestreo en el tiempo:
  - $2B$, con B ancho de banda efectivo del filtro de análisis $W(e^{j\omega})$
  - Hamming, Hanning: $B = \frac{2F_s}{L}$, Rectangular: $B = \frac{F_s}{L}$
  - Hamming: $L = 100$, $F_s = 10$ kHz, $B = 200$ Hz $\rightarrow$ cada 25 muestras
- tasa de muestreo en frecuencia:
  - L muestras en frecuencia para evitar aliasing temporal, $\omega_k = \frac{2\pi k}{L}$
- tasa de muestras total: $SR = 2BL$
  - para las ventanas típicas: $B = C_b \frac{F_s}{L} \rightarrow SR = 2C_b F_s$
  - $\frac{SR}{F_s} = 2C_b$, "sobremuestreo" de STFT respecto a $x[n]$
- en la práctica se pueden usar tasas más bajas

Detección de pitch usando STFT

- análisis del espectro de cada trama para estimar f_0
- ubicación del primer pico espectral ✗(no confiable, impreciso)
- todos los armónicos contribuyen (máximo común divisor)
- muchas propuestas para estimar f_0 a partir del espectro (ver [Hess, 2008])



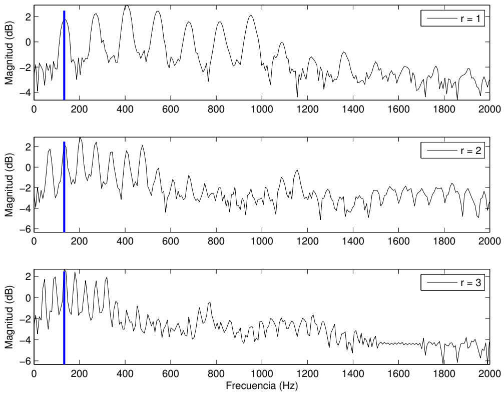
Detección de pitch usando STFT

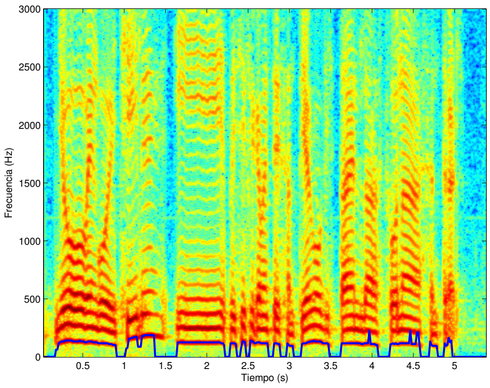
Producto armónico espectral

- producto de versiones comprimidas del espectro:

$$P_n(e^{j\omega}) = \prod_{r=1}^{K} |X_n(e^{j\omega r})|^2 \qquad \hat{P}_n(e^{j\omega}) = 2\sum_{r=1}^{K} \log |X_n(e^{j\omega r})|$$

- armónicos superiores refuerzan f_0, buena inmunidad al ruido





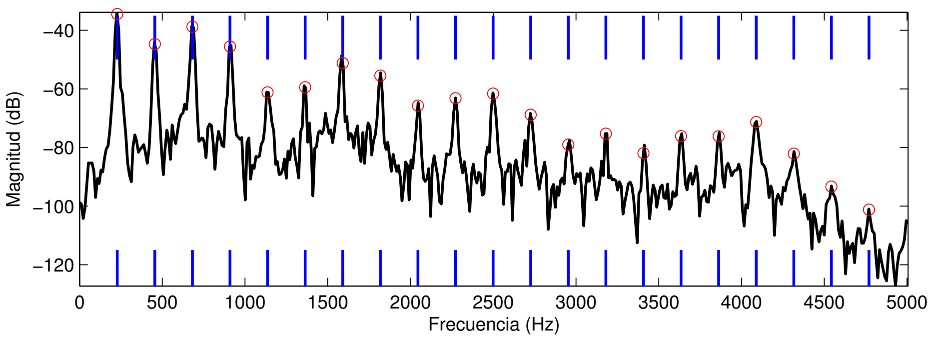
Detección de pitch usando STFT

Espectro logarítmico acumulado (GlogS, [Képesi and Weruaga, 2006])

- suma de la magnitud del espectro en posiciones armónicas

$$\rho_n(f_0) = \frac{1}{n_H} \sum_{i=1}^{n_h} \log |X_n(if_0)|$$

- el logaritmo se aplica para blanqueado del espectro
- grilla de valores de f_0, post-procesado para eliminar picos espúreos



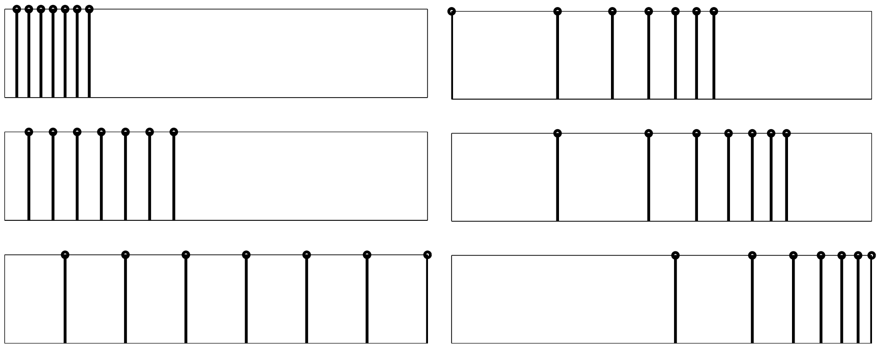
Detección de pitch usando STFT

Escala de frecuencia logarítmica

- posición relativa de componentes armónicos es constante

$$\begin{aligned} \log(2f_0) - \log(f_0) &= \log(2f_0/f_0) &= \log(2) \\ \log(3f_0) - \log(2f_0) &= \log(3f_0/2f_0) &= \log(3/2) \quad \ldots \end{aligned}$$

- posición absoluta del patrón depende de f_0
- detección de patrón para estimar f_0 [Brown, 1991]



Detección de pitch usando STFT

Escala de frecuencia: lineal vs. logarítmica

- DFT: resolución constante y escala lineal
  ejemplo: $N = 1024$ muestras, $F_s = 32kHz$
  - resolución: $\Delta f = F_s/N = 31.25Hz$
    resolución de semitono: $\sqrt[12]{2} = 1.0595$ i.e. 6%
  - violín, límite del registro bajo: $G_3 = 196$ Hz, resolución 16%
  - piano, límite del registro alto: $C_8 = 4186$ Hz, resolución 0.75%
- distribución exponencial permite resolución variable
  - semi-tono: 12 frecuencias por octava, $f_k = (2^{1/12})^k f_{min}$
  - cuarto-tono: 24 frecuencias por octava, $f_k = (2^{1/24})^k f_{min}$

  factor de calidad constante: $Q = f/\Delta f$
  - $Q = f_k/(f_{k+1} - f_k) = fk/(2^{1/12} - 1)f_k = 1/(0.0595) \approx 17$
  - $Q = f_k/(f_{k+1} - f_k) = fk/(2^{1/24} - 1)f_k = 1/(0.0293) \approx 34$

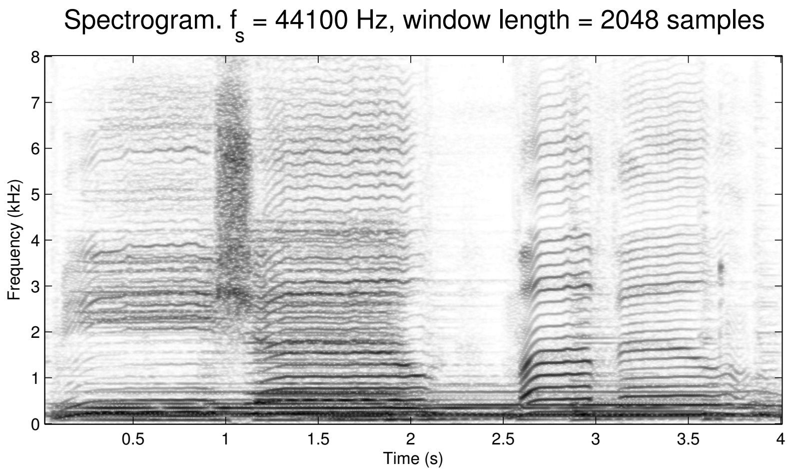
  resolución variable: $\Delta f = F_s/N_k$

  N_k: largo de ventana diferente para cada frecuencia

Análisis multiresolución

Características de señales de música

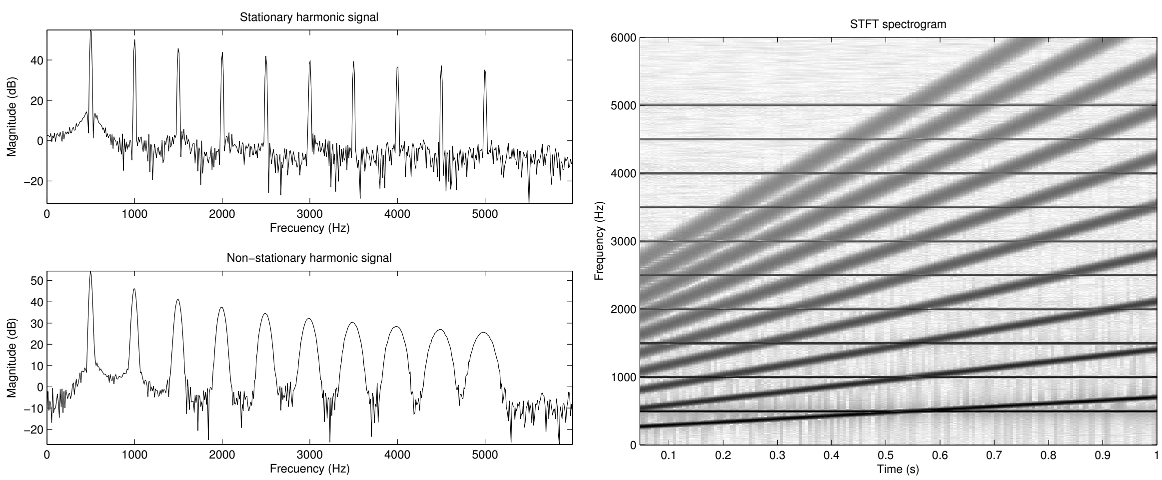
- sonidos de estructura armónica
- alta densidad de componentes en frecuencias baja y media
- modulación más notoria en alta frecuencia



Análisis multiresolución

Sonidos de estructura armónica

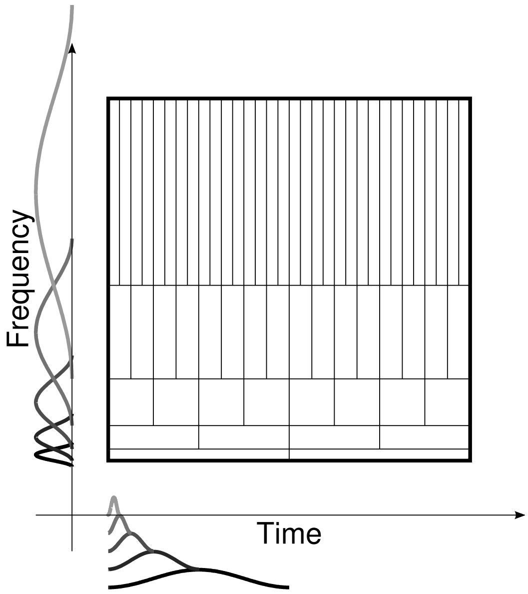
- buena aproximación de sonidos reales en intervalos de tiempo corto
- modulación en frecuencia $\rightarrow$ resolución pobre en alta frecuencia



Análisis multiresolución

Análisis multi-resolución

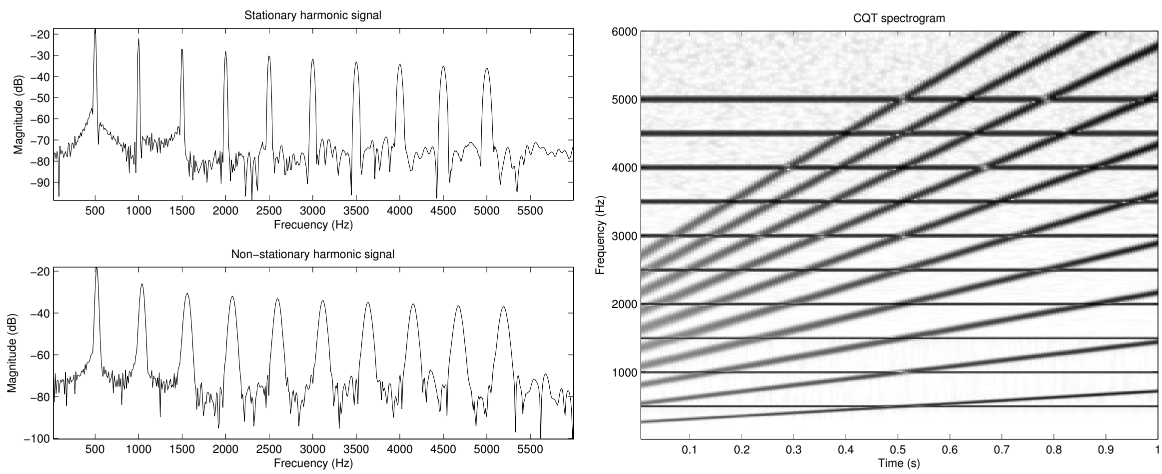
- cálculo de DFTs con diferente largo de ventana
- ejemplos:
  - Multi-resolution FFT (MRFFT) [Dressler, 2006]
  - Constant-Q Transform CQT [Brown, 1991]
  - Wavelets
- más apropiadas para el análisis de música



Análisis multiresolución

Comparación de STFT vs multi-resolución

- resolución tiempo-frecuencia mejorada para análisis multi-resolución

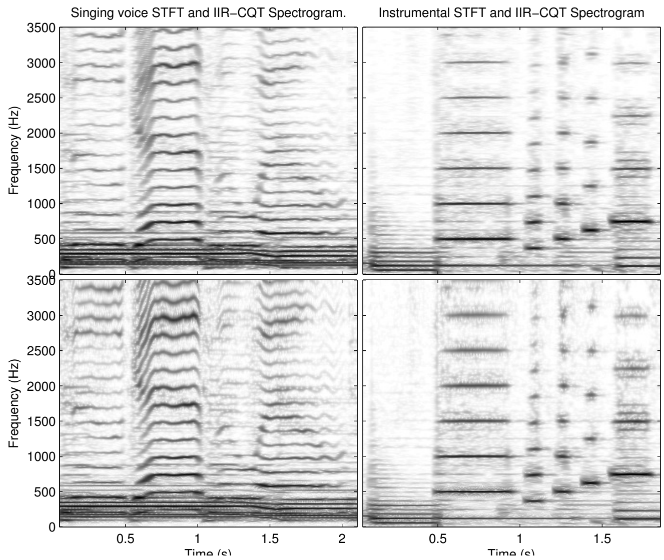


CQT: espectro y espectrograma

Análisis multiresolución

Comparación de STFT vs multi-resolución

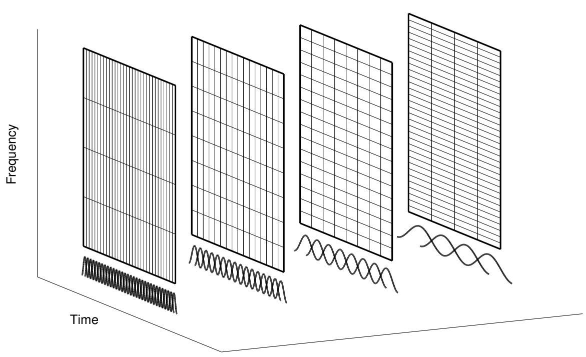
+ armónicos altos más precisos
+ más discriminación en baja frecuencia
+ inicio de notas mejor definido
− peor resolución sonidos estacionarios

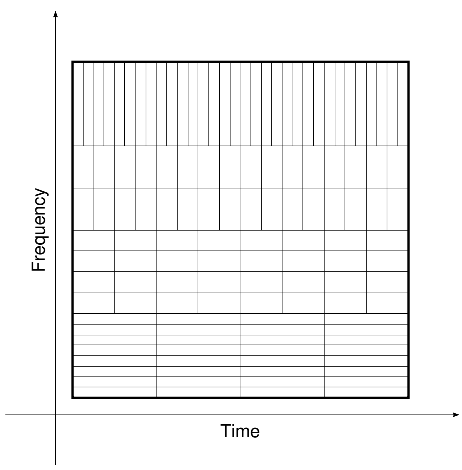


Análisis multiresolución

Multi-resolution FFT (MRFFT) [Dressler, 2006]

composición de STFT con diferentes largos de ventana
(e.g. $4096, 2048, 1024, 512$ muestras a $\mathrm{fs} = 44100$ Hz)



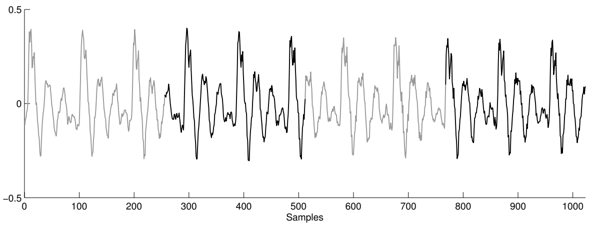


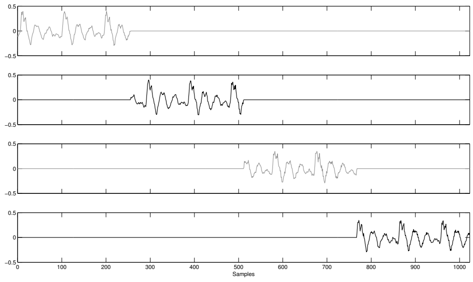
Análisis multiresolución

Multi-resolution FFT (MRFFT) [Dressler, 2006]

cálculo eficiente mediante suma de DFT de ventanas más cortas (transformadas elementales)

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-\frac{j2\pi kn}{N}} = \sum_{c=0}^{\frac{N}{L}-1} \sum_{n=cL}^{(c+1)L-1} x[n] e^{-\frac{j2\pi kn}{N}}.$$





e.g. DFT de 1024 muestras como suma de 4 DFT de 256 muestras

Análisis multiresolución

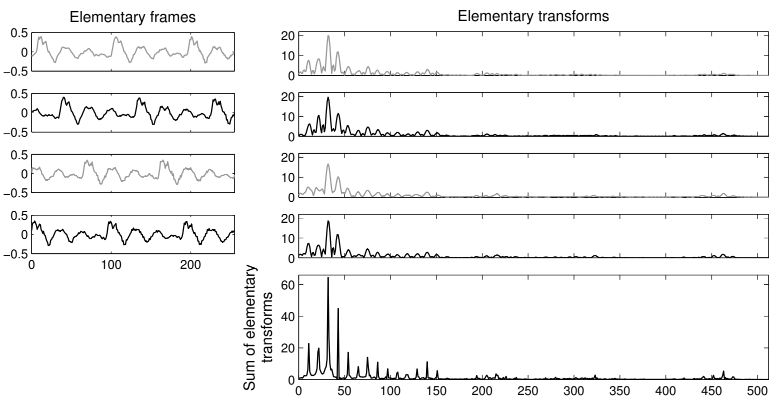
Multi-resolution FFT (MRFFT) [Dressler, 2006]

- corrimiento temporal efectuado en el dominio de la frecuencia para reutilizar transformadas elementales previas en tramas sucesivas

teorema de corrimiento de la DFT:

$$x[n] \overset{DFT}{\longleftrightarrow} X(k)$$

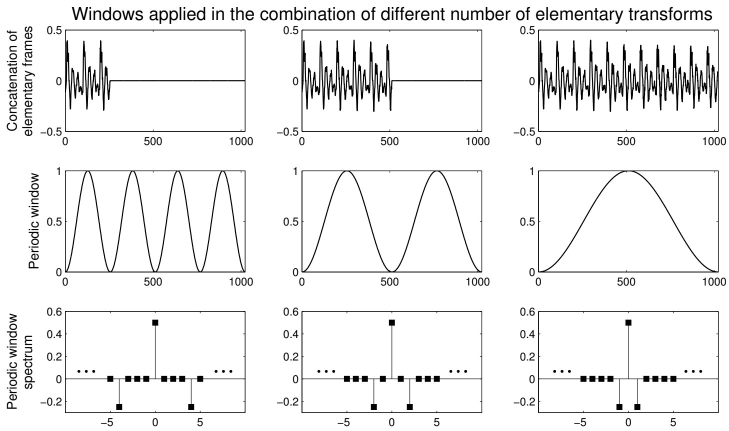
$$x[n+l] \overset{DFT}{\longleftrightarrow} X(k) e^{-j2\pi kl/N}$$



Análisis multiresolución

Multi-resolution FFT (MRFFT) [Dressler, 2006]

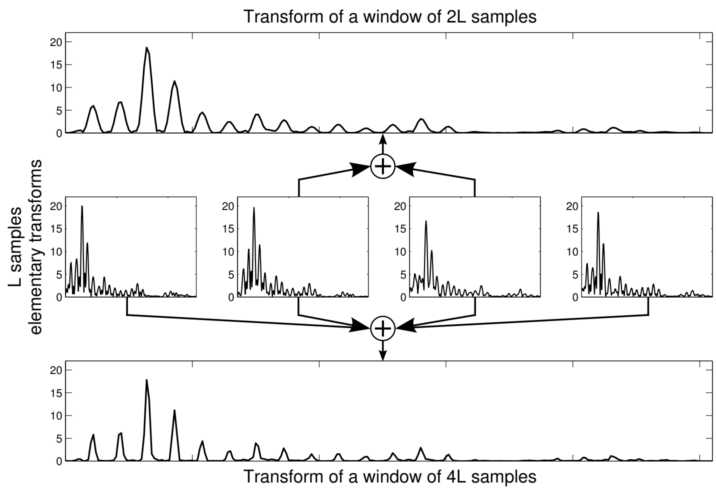
- enventanado mediante producto convolución en frecuencia
- ventanas con unos pocos componentes en frecuencia no nulos
- transformadas con agregado de ceros, se usan ventanas periódicas



Análisis multiresolución

Multi-resolution FFT (MRFFT) [Dressler, 2006]

ejemplo: usando transformadas elementales de 256 muestras
se combinan 8, 4 y 2 para obtener DFTs de 2048, 1024, y 512 muestras, usadas para representar frecuencias bajas, medias y altas respectivamente



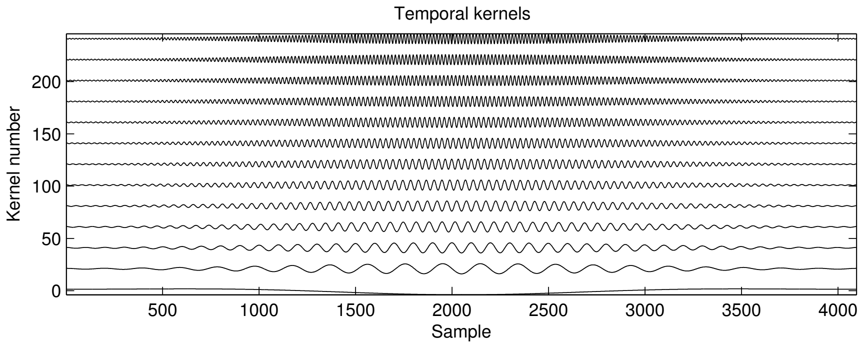
Análisis multiresolución

Constant Q Transform [Brown, 1991]

DFT:

$$X[k] = \sum_{n=0}^{N-1} w[n]x[n]e^{-j2\pi kn/N}$$

$$Q_k = f_k/\Delta f = k$$



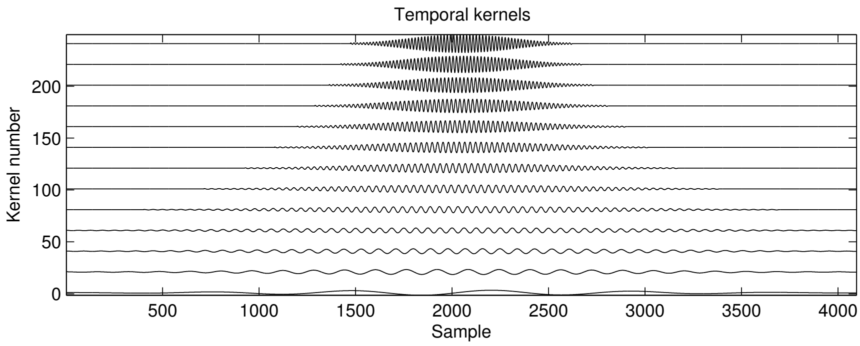
Análisis multiresolución

Constant Q Transform [Brown, 1991]

CQT:

$$\begin{aligned} Q &= f_k/\Delta f_k \text{ constante} \\ N_k &= Q/f_k \text{ variable} \end{aligned}$$

$$X^{cq}[k] = \frac{1}{N_k} \sum_{n=0}^{N_k-1} w_k[n]x[n]e^{-j2\pi Qn/N_k}$$



Análisis multiresolución

Constant Q Transform [Brown, 1991]

la evaluación directa de la CQT es computacionalmente costosa

aproximación eficientemente usando la FFT [Brown and Puckette, 1992]

- $X^{cq}[k] = \frac{1}{N_k}\sum_{n=0}^{N_k-1} w_k[n]x[n]e^{-j2\pi Qn/N_k}$
- $X^{cq} = x \cdot T^*$, multiplicación matricial con **kernels temporales**:

$$T^*[n,k] = \begin{cases} \frac{1}{N_k} w_k[n]e^{-j2\pi Qn/N_k} & \text{si } \ n < N_k \\ 0 & \text{en otro caso} \end{cases}$$

- usando la relación de Parseval para la DFT:

$$X^{cq}[k] = \sum_{n=0}^{N-1} x[n]T^*[n,k] = \frac{1}{N}\sum_{k'=0}^{N-1} X[k']K^*[k',k]$$

dónde $X[k']$ y $K[k',\cdot]$ son la DFT de $x[n]$ y $T[n,\cdot]$ respectivamente

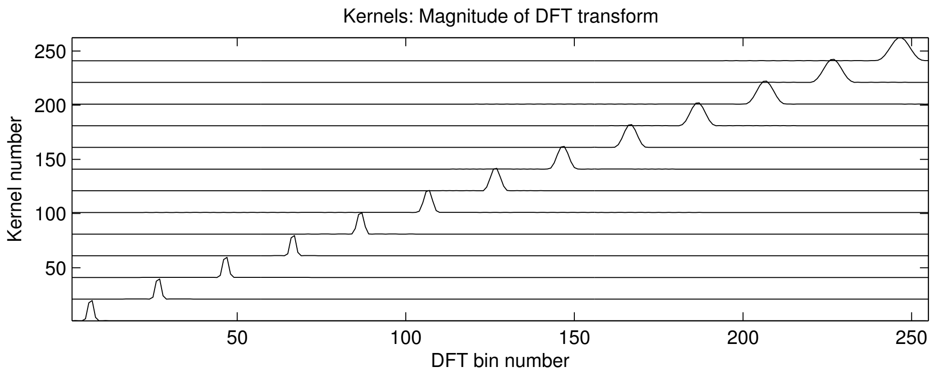
$K[k',\cdot]$ denominados **kernels espectrales**

Análisis multiresolución

Constant Q Transform [Brown, 1991]

en el caso de kernels temporales simétricos conjugados,

- los kernels espectrales son reales y cero para casi todo el espectro
- solo los componentes del kernel espectral superiores a un cierto umbral son considerados
- hay pocos productos involucrados en el cálculo de la CQT

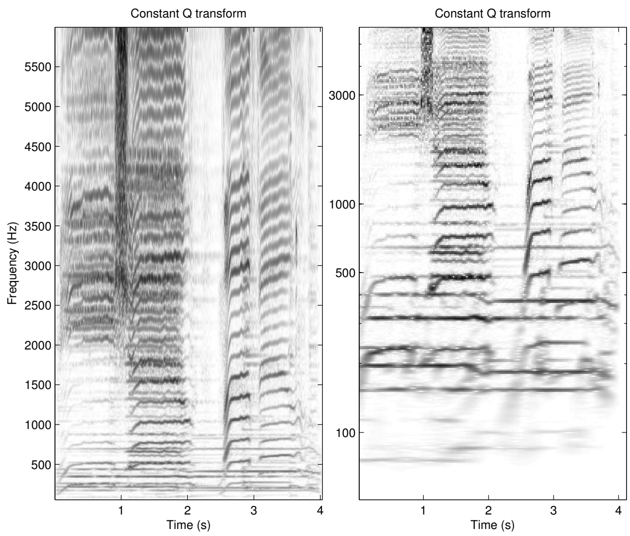


Análisis multiresolución

Constant Q Transform [Brown, 1991]

distribución de los bins en frecuencia

- la formulación original de la CQT implica distribución geométrica
- se puede formular para cualquier otro espaciado, por ejemplo lineal



Referencias I

Brown, J. C. (1991).
Calculation of a constant Q spectral transform.
JASA, 89(1):425–434.

Brown, J. C. and Puckette, M. S. (1992).
An efficient algorithm for the calculation of a constant Q transform.
JASA, 92(5):2698–2701.

Dressler, K. (2006).
Sinusoidal Extraction Using and Efficient Implementation of a Multi-Resolution FFT.
In *Proceedings of the DAFx-06*, Montreal, Canada.

Hess, W. (2008).
Pitch and voicing determination of speech with an extension toward music signals.
In *Springer Handbook of Speech Proc.*, pages 181–208. Springer, Heidelberg.

Referencias II

Képesi, M. and Weruaga, L. (2006).
Adaptive chirp-based time-frequency analysis of speech signals.
Speech Communication, 48(5):474–492.

Rabiner, L. R. and Schafer, R. W. (2011).
Theory and Applications of Digital Speech Processing.
Prentice Hall, 1st edition.
Chapter 7 - Frequency-domain representations.