

The Short-Time Fourier Transform

Xavier Serra
Music Technology Group
Universitat Pompeu Fabra, Barcelona

Index

- Introduction: the STFT
- Analysis window
- DFT computation
- Window hop-size
- Inverse STFT
- The Phase-vocoder

Introduction: STFT

$$X_l(k) = \sum_{n=0}^{N-1} w(n) x(n+lH) e^{-j\omega_k n} \quad l=0,1,\dots$$

w: real window,

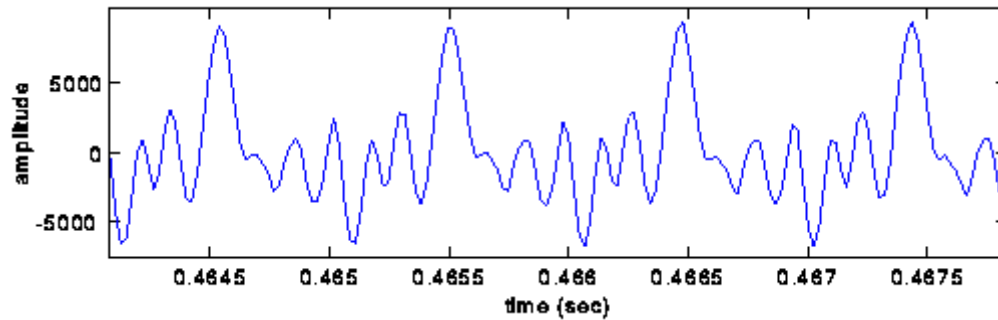
l: number of frame,

H: time advance of window (hop-size)

Interpretations:

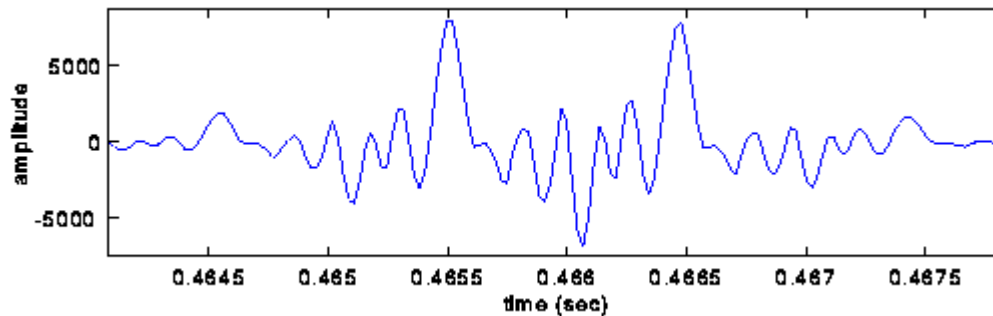
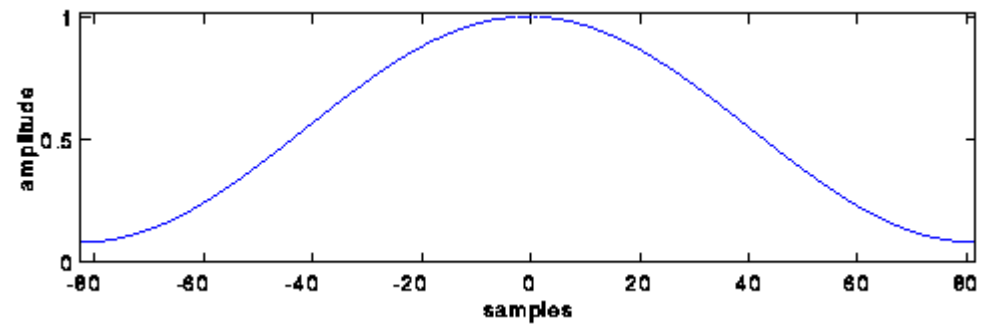
- overlap-add (l fixed)
- filter-bank (k fixed)

Analysis window



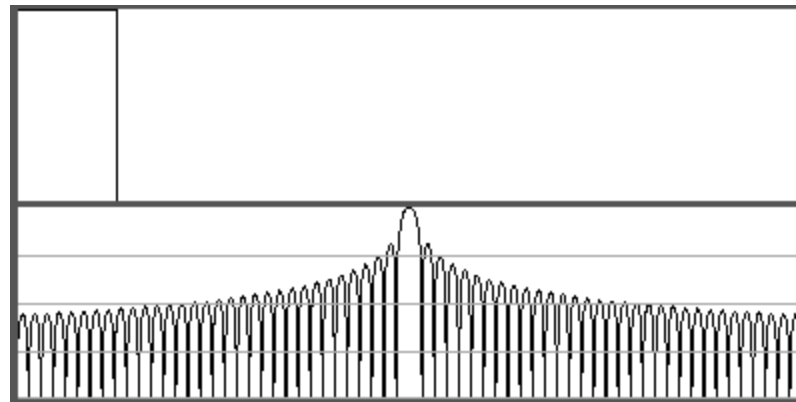
← $x(n)$

$w(n)$ →



← $y(n) = w(n)x(n),$
 $n = 0, 1, \dots, M-1$

- All standard windows are real and symmetric and have a frequency spectrum with a sinc-like shape.



rectangular
window

spectrum

- The choice is mainly determined by two of the spectrum's characteristics:
 1. Width of main lobe.
 2. Highest side-lobe level.

Window facts:

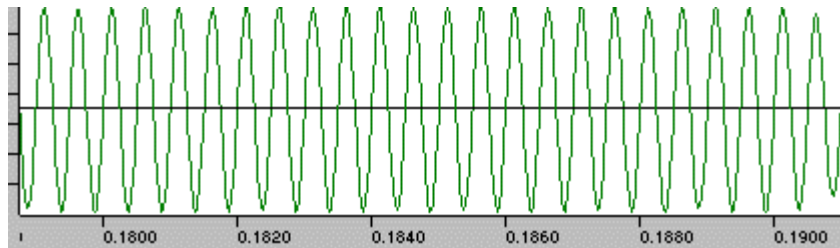
- It determines the trade-off of time versus frequency resolution.
- It affects the smoothness of the spectrum.
- It affects the detectability of different sinusoidal components.
- Most common windows are: Rectangular, Hanning, Hamming, Blackman-Harris, Kaiser.

Transform of a windowed sinusoid

complex sinusoid: $x(n) = Ae^{j\omega_x n}$, A : amplitude, ω_x : radian frequency

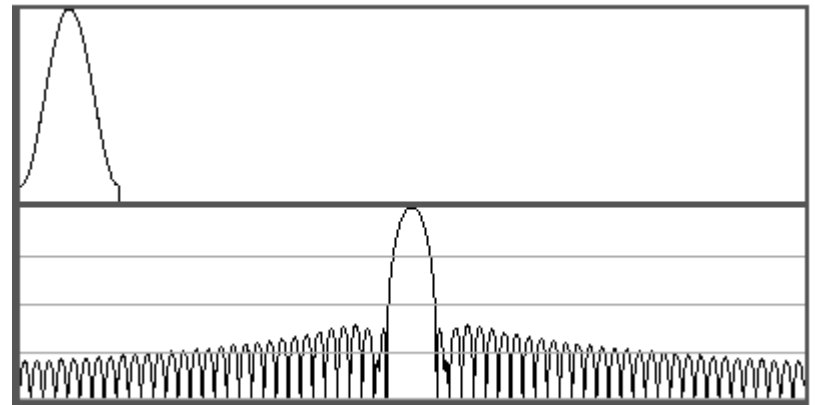
window: $w(n)$, $n=0,1,\dots,M-1$

$$\begin{aligned} X_w(\omega) &= \sum_{n=-\infty}^{\infty} x(n)w(n)e^{-j\omega n} \\ &= A \sum_{n=-M/2}^{(M/2)-1} w(n)e^{-j(\omega-\omega_x)n} \\ &= AW(\omega-\omega_x) \end{aligned}$$



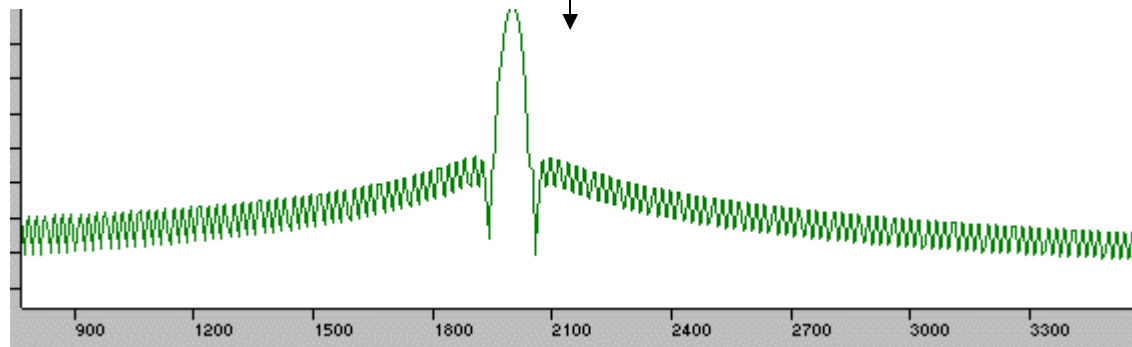
sine wave, 2,000 Hz

X



Hamming window

FFT



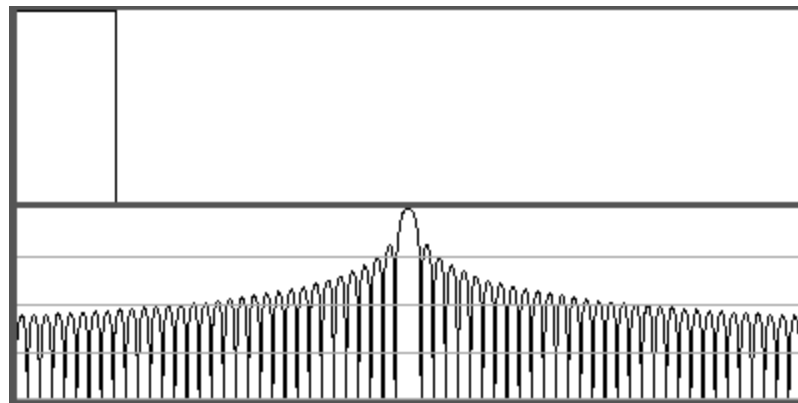
sine wave spectrum

Common windows

Rectangular:

$$w(n) = 1, \quad n = -M/2, \dots, 0, \dots, M/2$$
$$= 0, \quad \text{elsewhere}$$

$$W(\omega) = \frac{\sin(\omega M/2)}{\sin(\omega/2)}$$



main-lobe width: 2 bins
side-lobe level: -13 dB

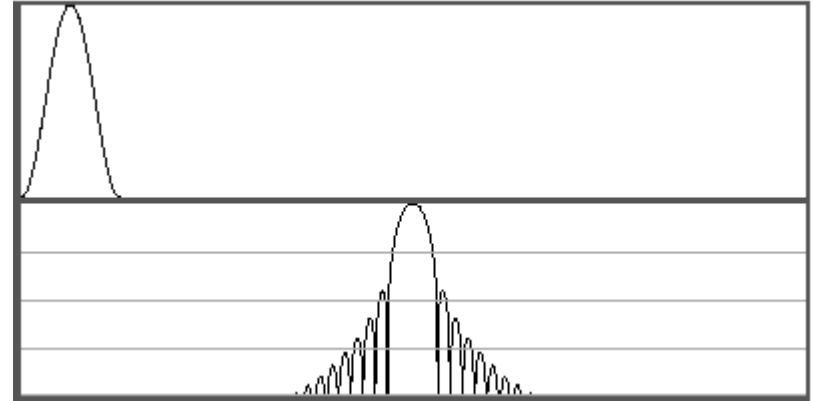
Hanning:

$$w(n) = .5 + .5 \cos(2n\pi/M),$$
$$n = -M/2, \dots, 0, \dots, M/2$$

$$W(\omega) = .5D(\omega) +$$
$$.25 \left[D\left(\omega - \frac{2\pi}{N}\right) + D\left(\omega + \frac{2\pi}{N}\right) \right]$$

$$\text{where } D(\omega) = \frac{\sin(\omega M/2)}{\sin(\omega/2)}$$

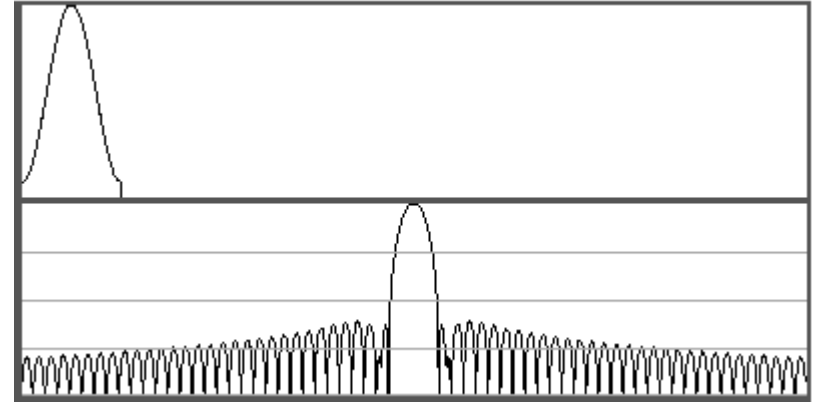
main-lobe width: 4 bins
side-lobe level: -23 dB



Hamming:

$$w(n) = 0.54 - 0.46 \cos(2n\pi/M), \\ n = 0, 1, \dots, M-1$$

$$W(\omega) = 0.54 D(\omega) - \\ 0.23 \left[D\left(\omega - \frac{2\pi}{N}\right) + D\left(\omega + \frac{2\pi}{N}\right) \right]$$



main-lobe width: 4 bins
side-lobe level: -43 dB

L-term Blackman-Harris:

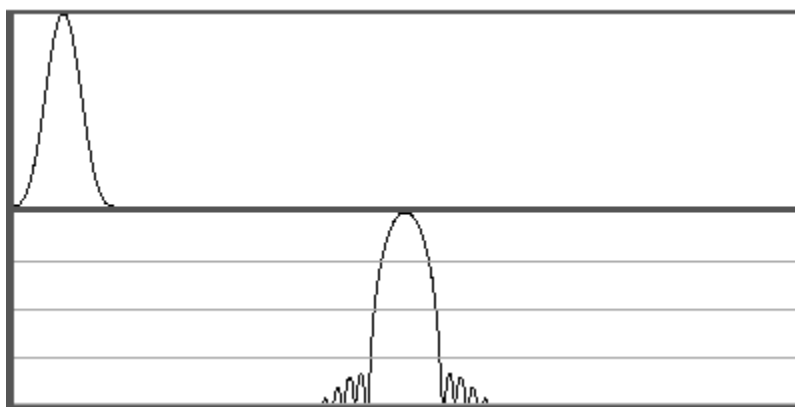
$$w(n) = \frac{1}{M} \sum_{l=0}^{L-1} \alpha_l \cos(2nl\pi/M), \quad n=0,1,\dots,M-1$$

Blackman-Harris 62dB: $\alpha_0=0.44859$ $\alpha_1=0.49364$ $\alpha_2=0.05677$

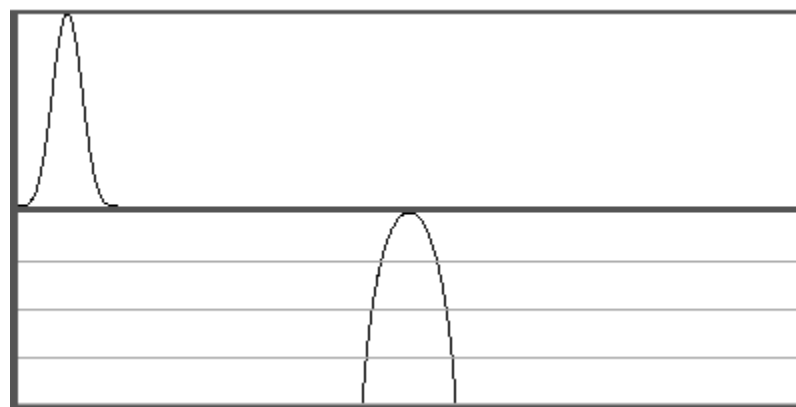
Blackman-Harris 70dB: $\alpha_0=0.42323$ $\alpha_1=0.49755$ $\alpha_2=0.07922$

Blackman-Harris 74dB: $\alpha_0=0.402217$ $\alpha_1=0.49703$ $\alpha_2=0.09892$ $\alpha_3=0.00188$

Blackman-Harris 92dB: $\alpha_0=0.35875$ $\alpha_1=0.48829$ $\alpha_2=0.14128$ $\alpha_3=0.01168$



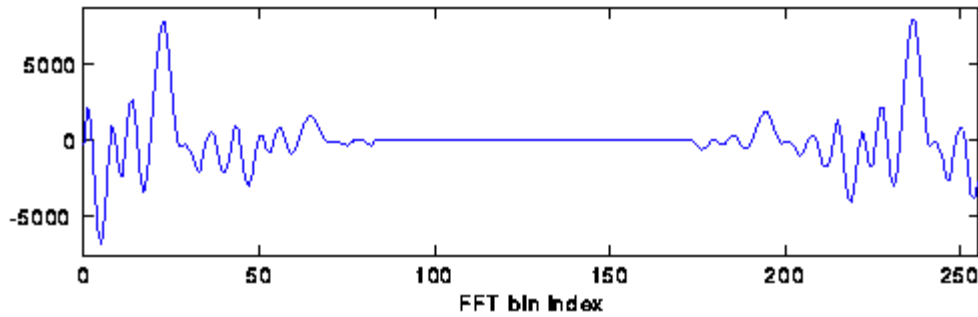
62 dB



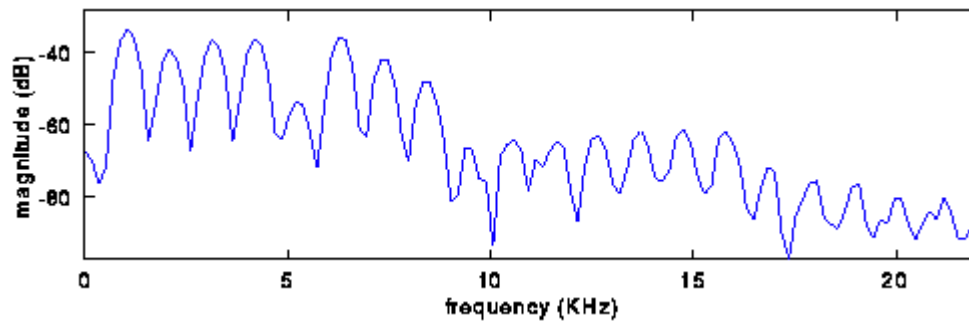
92 dB

DFT computation

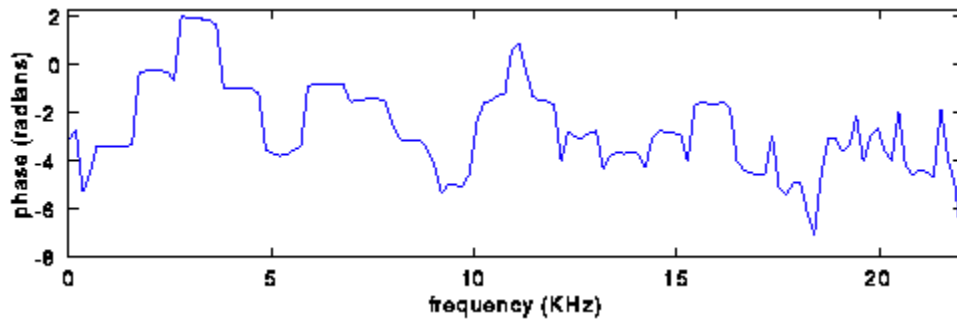
- **Window type:** chosen depending of the window size and time-frequency resolution desired.
- **Window size (M):** chosen depending of type of window and time-frequency resolution desired.
- **FFT size (N):** first power of two bigger than M . The difference ($N-M$) is filled with zeros (zero-padded)
- **Zero-phase window:** the windowed data (using an odd length window) is centered about the time origin.



FFT buffer



magnitude
spectrum



phase
spectrum

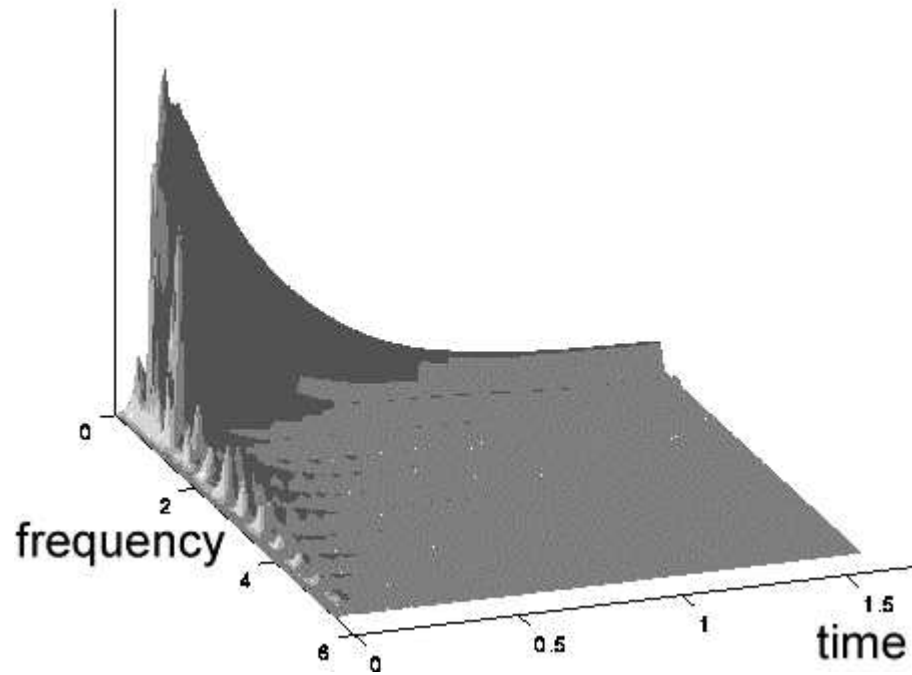
Window hop-size

- Successive frames should overlap in time in such a way that all the data are weighted equally.
- For certain windows exists perfect overlap factors.
Rectangular: M/j , Hanning and Hamming: $(M/2)/j$,
where $j = 0, 1, \dots$

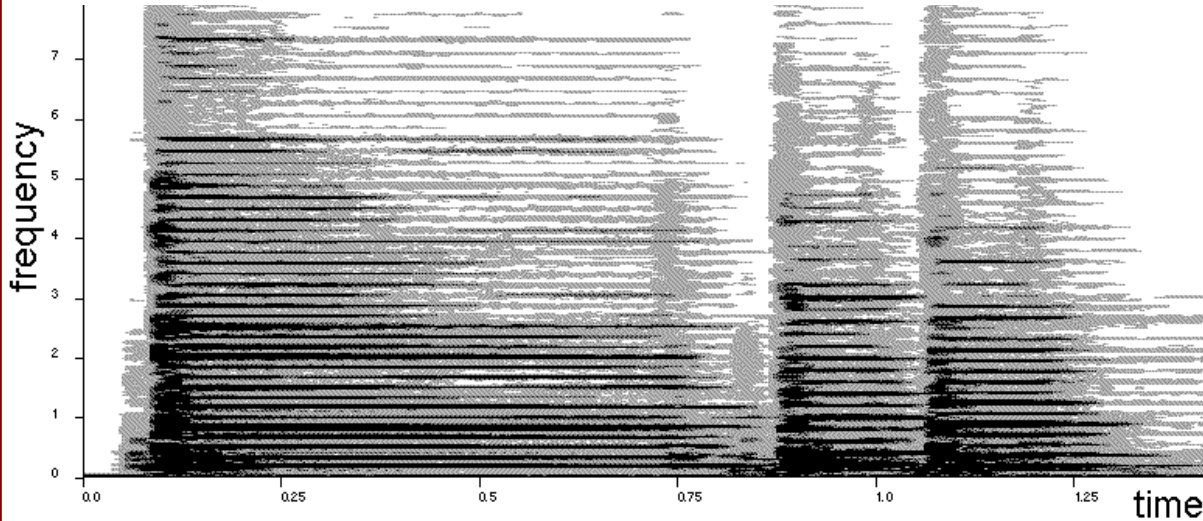
The overlap factor can be expressed by:

$$A_w(m) = \sum_{n=-\infty}^{\infty} w(m - nH) = c$$

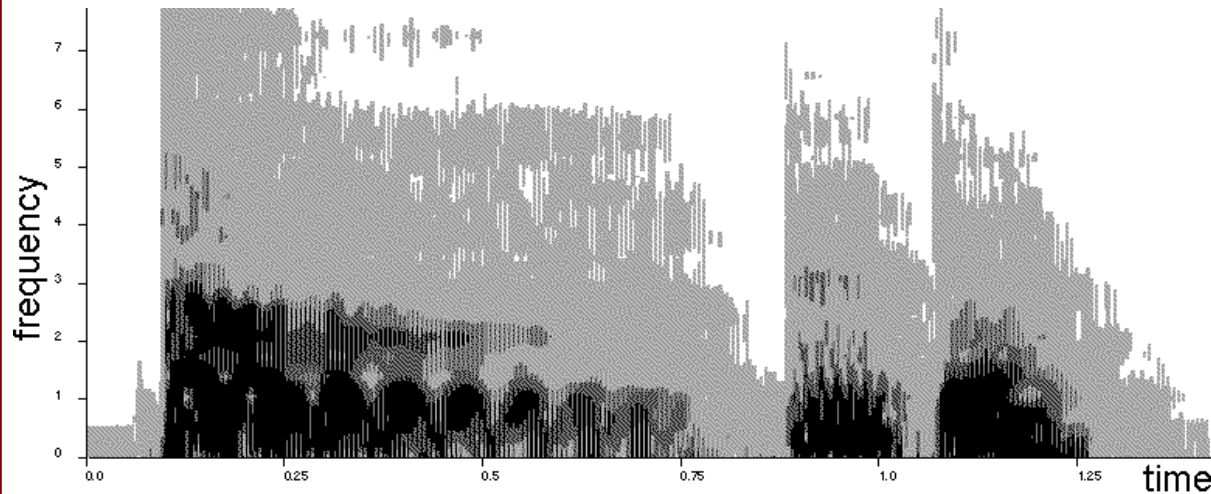
Spectrogram



Time-frequency compromise



good freq. resolution
bad time resolution



bad freq. resolution
good time resolution

Inverse STFT

The overlap-add interpretation of the STFT yields a particular synthesis method.

$$s(n) = \sum_{l=0}^{L-1} \text{Shift}_{lH, n} \left[\frac{1}{K} \sum_{k=0}^{K-1} X_l(k) e^{j\omega_k m} \right]$$

each synthesized frame is:

$$s_l(n) = \chi(n + lH) w(n)$$

and the synthesized sound is:

$$s(n) = \sum_{l=0}^{L-1} s_l(n - lH) = \chi(n) \sum_{l=0}^{L-1} w(n - lH)$$

The phase vocoder

Phase values are converted into instantaneous frequencies:

$$\hat{f}_l(k) = \frac{\theta_l(k) - \theta_{l-1}(k)}{2\pi HT}$$

H : hop-size of window,

$T (= 1/f_s)$: sample period,

\hat{f} : estimated frequency.

Phase is discarded and redefined as the integral of the frequency:

$$\hat{\theta}_m(k) = \hat{\theta}_{m-1}(k) + 2\pi T \hat{f}_m(k)$$

where $\hat{f}_m(k)$ is the linear interpolated frequency from $\hat{f}_{l-1}(k)$ to $\hat{f}_l(k)$