Analysis/resynthesis with the short time Fourier transform

summer 2006 lecture on analysis,
modeling and transformation of audio signals

Axel Röbel
Institute of communication science TU-Berlin
IRCAM Analysis/Synthesis Team

25th August 2006

Contents

1 The short time Fourier transform

- We have seen how a single analysis obtained with an analysis window cutting part of the signal can be used to investigate the **local properties** of the signal.
- The key idea for time frequency analysis is to use a sequence of such analysis frames to obtain a time varying spectral analysis of the signal.
- two types of graphical presentations are common: log amplitude and log energy.
- The result of such an analysis in log amplitude form is seen in fig. 1

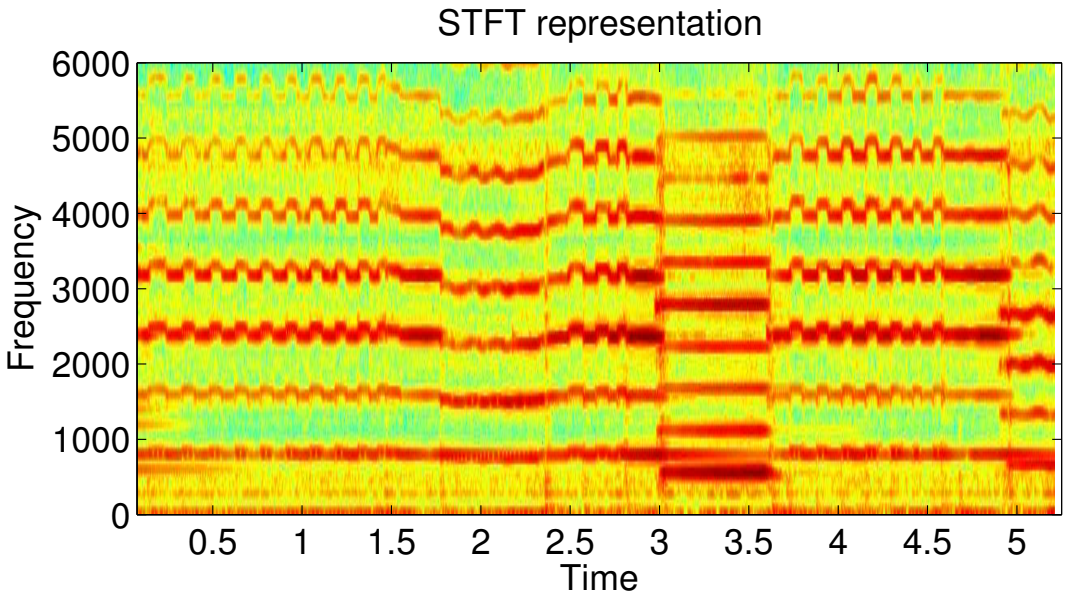

Figure 1: STFT amplitude spectrum of a musical performance

- observable features: pitch, vibrato, note onsets/offsets, voiced/unvoiced, volume, bandwidth, . . .

1.1 Interpretation as bank of filters

In a mathematical formulation the STFT is written as follows:

- sequence of DFT analysis with moving window w

$$X(n,k) = \sum_{m=n}^{N-1+n} w(m-n)x(m)e^{-j\Omega_N k(m-n)} \tag{1}$$

- for symmetric windows we may transform into

$$\begin{aligned} X(n,k) &= \sum_{m=n}^{N-1+n} w(m-n)x(m)e^{-j\Omega_N k(m-n)} && (2)\\ &= \sum_{m=-\infty}^{\infty} x(m)(w(n-m)e^{j\Omega_N k(n-m)}) && (3)\\ &= x(n) * ((w(n)e^{j\Omega_N kn}) = x(n) * h_w(n,k) && (4) \end{aligned}$$

- convolution of the signal with N bandpass filters at center frequencies $\Omega_N k$.
- each bandpass has
  - impulse response amplitude envelope corresponding to analysis window.
  - filter transfer function equal to amplitude spectrum of the window, centered at the respective frequency bin.
  - linear phase transfer function, due to the fact that the window is not centered at the origin.
- $X(n, k)$ represents the signal parameters for window position n and frequency $\frac{2\pi k}{N}$!
- as seen previously $X(n, k)$ represents sinusoidal values at the center of the window w that starts at time position n.
- **ATTENTION**: eq. (4) is used only for theoretical investigations!!! For practical applications the FFT analysis of the overlapping frames as in eq. (6) is generally much faster. While the gain depends on the STFT parameters, for common settings it is at least two orders of magnitude.

2 STFT parameters

The STFT parameters are window type and length L, FFT size N, frame offset (hop size) I.

- **MOST IMPORTANT!!** window size L is selected according to frequency and time resolution such that the interesting features (sinusoidal trajectories) are resolved.
- for harmonic sounds:
  $\rightarrow$ Mainlobe width $\approx$ fundamental frequency
  $\rightarrow L \approx \frac{4}{F_0}$
- what can be seen is determined mostly by the window size
- window type is selected according to compromise between sidelobe attenuation requirements and mainlobe width. (generally Hanning or Blackman)
- For best efficiency FFT size N is power of 2 larger than L (STFT) (or $2L$ sinusoidal parameter analysis).
- hop size according to bandwidth of bandpass outputs (see below).

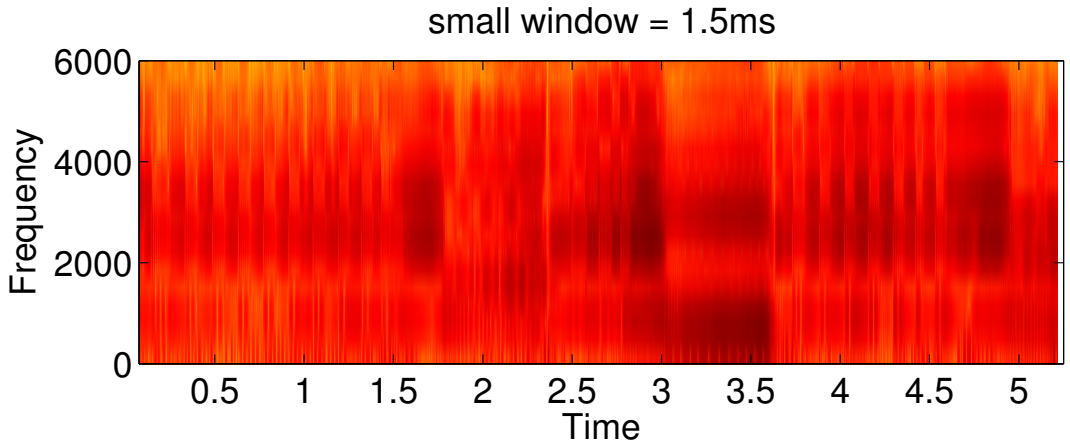

Figure 2: STFT amplitude spectrum with window size too small.

- Sinusoids are not resolved, we observe only some energy fluctuations
- weak relation to physical process

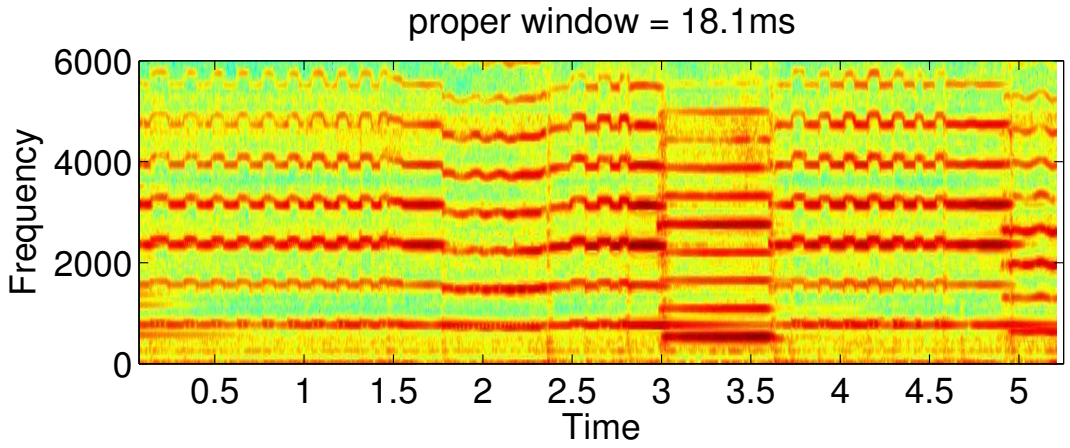

Figure 3: STFT amplitude spectrum with properly selected window sizes.

- all sinusoids are resolved,
- information about relevant physical parameters can be obtained

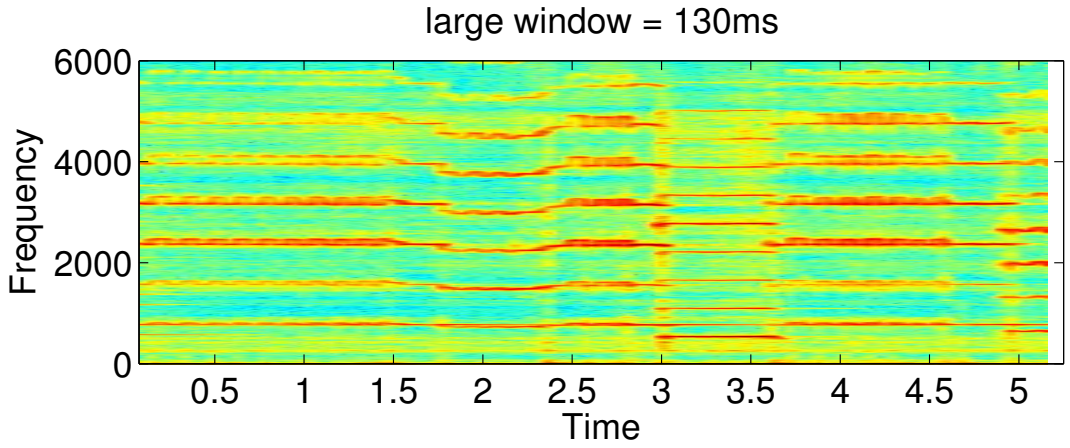

Figure 4: STFT amplitude spectrum window sizes too large.

- time variation of sinusoids cannot be observed correctly.
- sinusoids disconnect.
- information about relevant physical process cannot be obtained

Conclusion: Window size needs to be adapted to the signal and the information one is looking for.

2.1 Hop size

- Taking a frame offset of one we get for each bin of the STFT a complex signal with the same sample rate as the signal
- Because the bandwidth of the complex signal that represents the evolution of the spectrum over time is much smaller than the bandwidth of the original signal we have a rather redundant representation.
- using filter bank interpretation we conclude that the bandwidth of the signal attached to a particular bin is given by the mainlobe of the analysis window.
- we can calculate the mainlobe width B_M in [Hz] for a given analysis window and derive the frame offset according to the Shannon sampling theorem such that no aliasing occurs.

$$I = \frac{1}{B_w} \tag{5}$$

- for the rectangular window we get the required frame offset for alias free representation to $I \approx \frac{M}{2}$ for the Hanning window we get $I \approx \frac{M}{4}$ and for Blackman $I \approx \frac{M}{6}$

- mathematical formulation of sub-sampled STFT

$$X(lI, k) = \sum_{m=lI}^{lI+N-1} w(m - lI)x(m)e^{-j\Omega_N k(m-lI)} \quad (6)$$

$l \in \{0, \pm 1, \pm 2, \ldots\}$ is integer, I is hop size, N is FFT size, L is window size.

2.2 Invertibility

Can we construct the original signal from $X(lI, k)$?

- Inverse Fourier transform of one frame yields

$$w(n')x(n' + lI) = r_N(n') \sum_{k=0}^{N} X(lI, k) e^{j\Omega_N k n'} \tag{7}$$

- n' is coordinate system at the start position of the window
- $r_N(n')$ is rectangular window of length N which is used to select a single period of the periodic inverse transform
- transform coordinate system into signal time $n = n' + lI$

$$w(n - lI)x(n) = r_N(n - lI) \frac{1}{N} \sum_{k=0}^{N} X(lI, k) e^{j\Omega_N k(n-lI)} \tag{8}$$

- short cut notation for inverse DFT of individual STFT frame

$$x(lI, n) = w(n - lI)x(n) \tag{9}$$

- overlap add the contributions of all frames l using coordinate system of original signal

$$\frac{1}{N} \sum_{l=-\infty}^{\infty} (r_N(n - lI) \sum_{k=0}^{N} X(lI, k) e^{j\Omega_N k(n-lI)}) = \sum_{l=-\infty}^{\infty} x(lI, n) \tag{10}$$

$$= \sum_{l=-\infty}^{\infty} w(n - lI)x(n) \tag{11}$$

$$= x(n) \sum_{l=-\infty}^{\infty} w(n - lI) \tag{12}$$

- for each n for that

$$C(n) = \sum_{l=-\infty}^{\infty} w(n - lI) \neq 0 \tag{13}$$

we can divide by $C(n)$ to get

$$x(n) = \frac{1}{N} \frac{\sum_{l=-\infty}^{\infty} r_N(n - lI) \sum_{k=0}^{N} X(lI, k) e^{j\Omega_N k(n-lI)}}{\sum_{l=-\infty}^{\infty} w(n - lI)} \tag{14}$$

- if $N > L > I$ then $C(n) \neq 0$ and the input signal can be reconstructed from the STFT

3 Time-/Frequency domain modifications

3.1 Arbitrary modifications

- We may apply arbitrary operators to modify a given STFT of a signal
- Due to the overlap of the analysis windows the STFT of a signal obeys a characteristic correlation between the neighboring frames.
- If these correlations are not respected, there is no signal that corresponds with the given STFT.
- The question to be solved: how do we generate a signal that has a STFT as close as possible to the STFT at hand.
- looking for minimum squared error solution

  the target STFT is $Y(lI, k)$, and we are looking for the signal $x(n)$ such that

$$\sum_{l=-\infty}^{\infty} \sum_{k=0}^{N-1} |X(lI, k) - Y(lI, k)|^2 = \quad \text{MIN} \tag{15}$$

- the solution (see section **4.1**) has been given by Griffin and Lim in 1984.

$$x(n) = \frac{\sum_{l=-\infty}^{\infty} w(n - lI)y(lI, n)}{\sum_{l=-\infty}^{\infty} w(n - lI)^2} \tag{16}$$

  where $y(lI, n)$ is the inverse DFT of each individual frame $Y(lI, k)$ and $w(n)$ is the analysis window.

- procedure to obtain inverse STFT of modified or unmodified STFT:
  1. inverse transform individual frames
  2. apply analysis window as synthesis window to inverse transform
  3. overlap add and keep track of the squared sum of synthesis window factors applied to each time position n
  4. normalize.
- procedure yields original signal if STFT has not been modified and signal with optimally close STFT in case the STFT has been modified.

4 Appendix

4.1 Reconstruction from modified STFT

We assume that we have a given STFT $Y(lI, k)$ that has DFT size N, hop size I and analysis window w. Y may be modified such that there exists no signal $x(n)$ with a STFT $X(lI, k) = Y(lI, k)$. Nevertheless we may the signal $x(n)$ that, if transformed into the STFT domain will minimize the squared error with respect to $Y(lI, k)$

$$D = \sum_{l=-\infty}^{\infty} \sum_{k=0}^{N-1} |X(lI, k) - Y(lI, k)|^2 = \quad \text{MIN} \tag{17}$$

According to the Linearity of the DFT and to Parseval's relation for the DFT we may replace the inner sum by means of the sum over the corresponding time signal

$$D = \frac{1}{N} \sum_{l=-\infty}^{\infty} \sum_{k=0}^{N-1} |x(lI, n) - y(lI, n)|^2 \tag{18}$$

In this equation we are given $y(lI, n)$ and we are searching for $x(n)$ that is related to

$x(lI, n)$ according to

$$D = \frac{1}{N} \sum_{l=-\infty}^{\infty} \sum_{n=0}^{N-1} |w(n - lI)x(n) - y(lI, n)|^2 \tag{19}$$

Calculating the derivative with respect to $x(n)$ and setting to zero we obtain the result of Griffin and Lim, eq. (16), as follows

$$0 = \sum_{l=-\infty}^{\infty} 2(w(n - lI)x(n) - y(lI, n))w(n - lI) \tag{20}$$

$$x(n) \sum_{l=-\infty}^{\infty} w(n - lI)^2 = \sum_{l=-\infty}^{\infty} y(lI, n)w(n - lI) \tag{21}$$

$$x(n) = \frac{\sum_{l=-\infty}^{\infty} y(lI, n)w(n - lI)}{\sum_{l=-\infty}^{\infty} w(n - lI).^2} \tag{22}$$

(23)

which shows that the optimal signal is obtained by means of normalization with the squared window.

References